ActFormer: Scalable Collaborative Perception via Active Queries

Suozhi Huang[1,2,*], Juexiao Zhang[1,*], Yiming Li[1,✉], and Chen Feng[1,✉]

https://coperception.github.io/ActFormer/

***Abstract*— Collaborative perception leverages rich visual observations from multiple robots to extend a single robot's perception ability beyond its field of view. Many prior works receive messages broadcast from all collaborators, leading to a scalability challenge when dealing with a large number of robots and sensors. In this work, we aim to address *scalable camera-based collaborative perception* with a Transformer-based architecture. Our key idea is to enable a single robot to intelligently discern the relevance of the collaborators and their associated cameras according to a learned spatial prior. This proactive understanding of the visual features' relevance does not require the transmission of the features themselves, enhancing both communication and computation efficiency. Specifically, we present ActFormer, a Transformer that learns bird's eye view (BEV) representations by using predefined BEV queries to interact with multi-robot multi-camera inputs. Each BEV query can actively select relevant cameras for information aggregation based on pose information, instead of interacting with all cameras indiscriminately. Experiments on the V2X-Sim dataset demonstrate that ActFormer improves the detection performance from 29.89% to 45.15% in terms of AP@0.7 with about 50% fewer queries, showcasing the effectiveness of ActFormer in multi-agent collaborative 3D object detection.**

I. Introduction

Collaborative perception, such as collaborative object detection [1] and semantic segmentation [2], empowers autonomous robots to share their perceptual insights, fostering a comprehensive understanding of their surrounding environments. It addresses challenges such as occlusions and sparse sensory information over long distances, which often impede individual perception. Nevertheless, scalability poses a notable challenge to current learning-based collaborative perception methods. This challenge mainly stems from the passive nature of existing approaches, which incorporate all accessible sensor data at some point, rather than *actively* requesting only the essential information before initiating any form of communication with collaborators.

Certain prior approaches [3, 4, 5] involve selective data usage to minimize communication overhead. However, the decision-making process for selecting sensor data still relies on transferred feature representations of the sensor measurements from other collaborators. As a result, these approaches are not truly *active* and struggle to scale up effectively when dealing with a large number of sensors or robots. Hence, this work aims to address the problem of *scalable collaborative 3D perception based on 2D camera input*. We propose to enable the ego-robot to *actively and intelligently request and use information*, as opposed to *passive and indiscriminate utilization of all available camera images*.

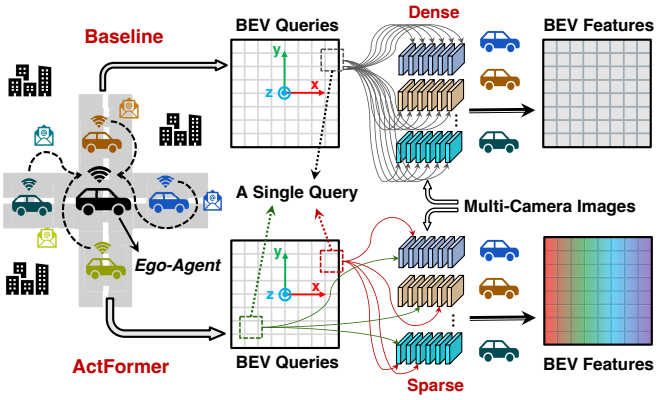


Fig. 1. **Baseline v.s. ActFormer.** Baseline densely attends BEV queries to every 2D image feature. ActFormer actively selects queries based on spatial information and achieves more scalable collaborative perception.

Inspired by the recent advances of Transformer for camera-based single-robot perception [6, 7], we design our method based on this query-based framework that leverages a learnable grid-shaped bird's eye view (BEV) queries corresponding to the target 3D space, to interact with the 2D camera features for 3D representation learning. Yet existing query-based methods like BEVFormer [7] are not designed to handle scalable camera-based collaborative perception. Therefore, each BEV query is supposed to interact with multi-agent multi-camera input in a dense manner, severely limiting the scalability and efficiency of the collaborative perception system, as shown in Fig. 1 (top branch).

In practice, querying all available sensory streams densely is highly redundant and inefficient. Our key idea is to optimize the query graph by enabling each BEV query to identify the most relevant 2D camera features based on the poses of robots and their cameras. To implement this idea effectively, we employ a learnable active selection module. This module takes input in the form of the ego robot's BEV features and the poses of collaborators, generating an interest score for each BEV query with respect to the available cameras. Only the high-score queries will be involved in the computation during the collaboration. Our method, termed *ActFormer*, creates a sparse query graph for extracting a

* indicates equal contributions. Work done during Suozhi's visit at NYU.
✉ Corresponding authors. This work is supported by NSF Grant 2238968. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

[1]Juexiao Zhang, Yiming Li, and Chen Feng are with New York University, New York, NY 11201, USA {juexiao.zhang, yimingli, cfeng}@nyu.edu

[2]Suozhi Huang is with Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University, Beijing 100084, China huang-sz20@mails.tsinghua.edu.cn

robust BEV representation from multi-robot multi-camera input, as illustrated in Fig. 1 (bottom branch), largely boosting both efficiency and scalability for collaborative perception. We test ActFormer in the task of collaborative 3D object detection from 2D images and conduct comprehensive experiments on the widely-used V2X-Sim [8] dataset. Quantitative and qualitative results show that ActFormer not only boosts the detection performance by a large margin compared to the baseline method ($45.88\% \rightarrow 53.33\%$ for AP@0.5; $29.89\% \rightarrow 45.15\%$ for AP@0.7), but also reduces the computational cost with 50% fewer queries. In summary, our major contributions are summarized as follows:

- We conceptualize a scalable and efficient collaborative perception framework that can actively and intelligently identify the most relevant sensory measurements based on spatial knowledge, without relying on the sensory measurements themselves.
- We ground the concept of the scalable collaborative perception with a Transformer, *i.e.*, *ActFormer*, which uses a group of 3D-to-2D BEV queries to actively and efficiently aggregate the features from multi-robot multi-camera input, only relying on pose information.
- We conduct comprehensive experiments in the task of collaborative object detection to verify the effectiveness and efficiency of our ActFormer.

II. Related Works

A. Collaborative perception

Collaborative perception has been shown to enhance the adaptability, robustness and efficiency of individual autonomous vehicle systems [2, 5, 9, 10, 11, 12, 13, 14]. Many of them deploy feature-level collaborative perception techniques, where the intermediate features produced by the deep neural networks are shared among a team of robots, be it as a swarm of drones [3, 15] or a fleet of autonomous vehicles [16, 17, 18]. Due to the high dimensionality of the intermediate features, collaboration cost is a critical issue to be considered. When2com [3] applies an attention mechanism over all the potential collaborators to fuse their features. Who2com [4] deploys a multi-stage handshake mechanism where the ego agent determines who to connect with based on the first stage of information exchange. V2VNet [16] trains a graph neural network for information aggregating. DiscoNet [1] learns about message selection via knowledge distillation [19] from an oracle model that is trained with the complete scene observation. STAR [17] learns to amortize communicating the intermediate representations over a few consecutive time steps with a spatio-temporal masked autoencoder [20].

Where2comm [5] shares with us a similar insight to use spatial information to guide collaboration, but its confidence map is based on LiDAR encodings and it still requires obtaining the feature representations of the sensor measurements from collaborator robots. Meanwhile, most methods focus on the input of 3D LiDAR point clouds instead of 2D camera images. CoCa3D [12] tackles camera-based detection with the help of depth estimation. However, they did not address the scalability and efficiency issues. In a word, it is still underexplored how to actively request and exploit the sensor information of neighboring robots for scalable collaborative 3D perception from 2D camera images, without relying on the sensor measurements themselves.

B. Transformer-based camera-only 3d perception

3D perception tasks have been central to autonomous driving research [21, 22, 23, 24, 25, 26]. Recently, Transformers [27] are successfully adopted into camera-only 3D perception tasks, including but not limited to 3D object detection [6], semantic scene completion [28], *etc*. Many of these Transformer-based models achieve superior performance by modeling an intermediate BEV representation before producing the task-specific outputs. [29, 30, 31] estimate the depth distribution then project the 2D image features to 3D space to get the BEV voxel features. Some treat the grid-shaped BEV features as queries or proposals and apply attention mechanisms over them and input features [32, 33]. Their BEV features are directly associated with the task outputs such as the object bounding boxes or voxel labels.

However, when involved with BEV features, the multi-head attention mechanism in the original Transformer can induce expensive computational costs since the number of queries is likely to be large. To overcome such computational bottleneck, deformable attention [34] is proposed which changes the original attention head from attending every token to only some of the tokens by learnable offsets:

$$\mathrm{DfmAttn}(q,p,F)=\sum_{m=1}^{N_{\text{head}}}\mathcal{W}_m\sum_{n=1}^{N_{\text{key}}}\mathcal{A}_{mn}\cdot\mathcal{W}'_m F\left(p+\Delta p_{mn}\right) \tag{1}$$

where q, p, F represent the query, reference point, and visual features, respectively. N_{head} denotes the number of attention heads and N_{key} is the number of sampled keys per attention head. $\mathcal{W}_m \in \mathbb{R}^{C\times(C/H_{\text{head}})}$ and $\mathcal{W}'_m \in \mathbb{R}^{(C/H_{\text{head}})\times C}$ are the learnable weights, where C is the feature dimension. $A_{mn} \in [0,1]$ is the normalized attention weight $\sum_{n=1}^{N_{\text{key}}} A_{mn} = 1$. $\Delta p_{mn} \in \mathbb{R}^2$ are the predicted offsets to the reference point p. $F\left(p+\Delta p_{mn}\right)$ represents the feature at location $p+\Delta p_{mn}$ in an image, extracted by bilinear interpolation as in [35]. It is adopted from [36] and extended to BEV representation learning in the BEVFormer [7], which is the state-of-the-art method in camera-only 3D object detection.

In BEVFormer [7], 3D query points are projected from 3D BEV space to 2D image space. Denote the BEV queries set as grid-shape $Q \in \mathbb{R}^{H\times W\times C}$ then $q \in \mathbb{R}^{1\times C}$ represents the query at a spatial location $(x,y) \in \mathbb{R}^{H\times W}$, denoted as $q = Q(x,y)$. Then the spatial cross attention between the query q and image features:

$$\mathrm{SCA}\left(q,F\right)=\frac{1}{\left|\mathcal{R}_{\text{hit}}\right|}\sum_{i\in\mathcal{R}_{\text{hit}}}\sum_{j=1}^{N_{\text{ref}}}\mathrm{DfmAttn}\left(q,\mathcal{P}(q,i,j),F_i\right) \tag{2}$$

note that only the images whose field of view is hit by query q, denoted as the set $\mathcal{R}_{\text{hit}}$, are considered. For each BEV

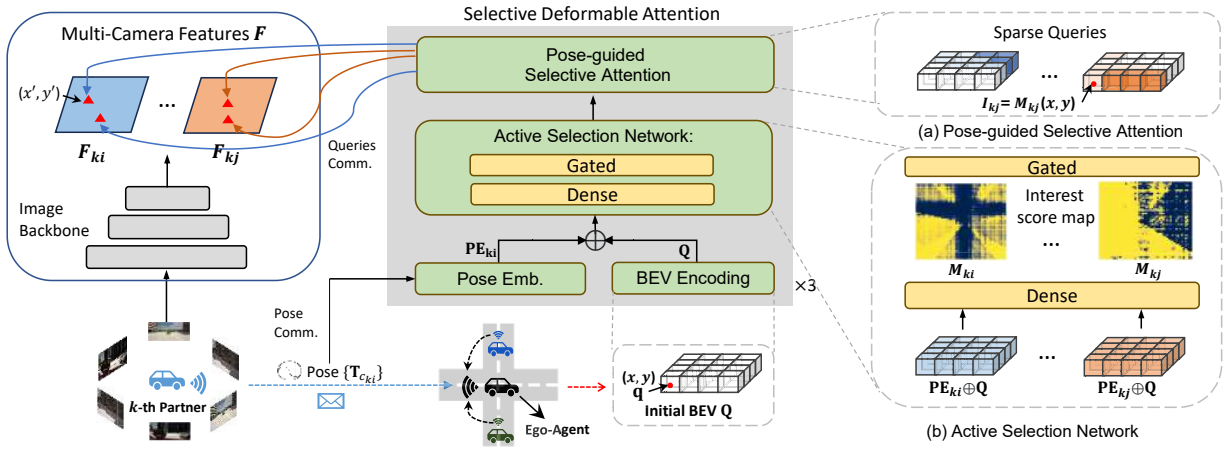


Fig. 2. **Method overview.** After partners broadcast their pose information to the ego car, our approach leverages selective deformable attention to obtain active sparse queries for images. Selective deformable attention consists of two crucial components: (a) **Pose-guided Selective Attention**, which efficiently focuses on multi-agent image features using active queries, enabling us to emphasize regions of interest; and (b) **Active Selection Network**, which concatenates pose embeddings with BEV queries and produces an interest score map. Subsequently, this interest score map is multiplied by the BEV query using a gated network to obtain the active query. This process aims to enhance collaboration efficiency and generate active sparse queries.

query q, project function $\mathcal{P}(q,i,j)$ gets the j-th reference point on the i-th view image by transforming with matrix $\mathbf{T_i}$. And N_{ref} indicates the number of reference points along a z-axis on each BEV query position. Hence for each query q, a pillar of 3D reference points $(x,y,z_j)_{j=1}^{N_{\text{ref}}}$ are projected to different image views through the specific projection matrix of cameras, which can be written as:

$$p = \mathcal{P}_k(q,i,j) = \mathbf{T_i}(x,y,z_j)^\intercal = \left(x'_{ij}, y'_{ij}\right) \ \forall q,i,j \quad (3)$$

x'_{ij}, y'_{ij} are in the image space and p is the 2D reference point as in Eq. 1. Our ActFormer is built based on deformable attention. We extend the above SCA with an active selection mechanism that further selects queries based on the current vehicles' poses and enables scalable and efficient multi-agent collaboration.

III. Method

Since the task focuses on autonomous driving, we refer to the collaborating robots as *vehicles* for clarity. Specifically, during the collaboration, any vehicle, referred to as the *ego* vehicle when considering its perception, can actively select information from other vehicles, referred to as the *partners*, to attend to, based on its current Bird's Eye View (BEV) representation and the relative poses of the others. Then the ego vehicle collaborates on active BEV queries with the corresponding partners' features and updates its own BEV representation accordingly. Finally, it decodes the detection output based on the updated BEV representation.

A. Motivation

Imagine a scenario where an ego vehicle encounters a complex urban intersection with multiple partner vehicles in the same scene. Each vehicle is equipped with multiple cameras placed at various poses. Collaborative perception enables these vehicles to share their observations and achieve a higher level of perception than they could individually. Our motivation stems from the idea that how vehicles collaboratively perceive should be closely related to their relative poses. Different camera poses result in varying viewpoints, each capturing unique information. However, conventional collaborative methods often treat all viewpoints equally, overlooking the fact that these camera perspectives offer different insights into the environment—some unique, some overlapping, and some redundant. Consequently, the ego vehicle may not fully capitalize on the diverse perspectives available, leading to indiscriminate collaboration that generates excessive communication and computation. Actually, communication may not be necessary when some partners share very similar observations.

Therefore, we believe that leveraging the poses can effectively guide collaboration. When a vehicle is provided with its partners' poses and their cameras' poses, it can estimate and assign varying degrees of significance to different camera viewpoints based on their spatial relationship and relevance to the ego's perception. In doing so, the ego vehicle can make informed and proactive selections about which sensory inputs to prioritize, thereby enhancing its perception capacity while reducing communication redundancy. This active selection mechanism recognizes the importance of the ego's self-awareness and adaptability and empowers ego vehicles to make contextually relevant decisions. This not only improves perception but also fosters a more dynamic, scalable, and efficient collaborative environment for autonomous robots. We will elaborate on the details in the following sections.

B. Definition

Consider N agents in total each with M cameras in the scene. Let $\mathcal{B}$ be the ego agent's BEV feature encoding.

Let $\{\mathcal{X}_i\}_{i=1}^{M}$ and $\mathcal{Y}$ be the observation sets of all cameras and the perception supervision of ego agent, respectively. The objective of collaborative perception is to achieve the maximized perception performance of all agents as a function of the number of selected agents N; that is,

$$\xi_{\Phi}(N) = \arg\max_{\theta,\mathcal{P}} \sum_{k=1}^{N} g\left(\Phi_\theta\left(\mathcal{B}\left(\{\{\mathcal{X}_i\}_{i=1}^{M}\}_{k=1}^{N}\right)\right), \mathcal{Y}\right) \tag{4}$$

where $g(\cdot,\cdot)$ is the perception task, specifically it can be 3D object detection evaluation metrics such as bounding boxes for mAP evaluation. Φ is the perception network with trainable parameter θ, and $\{\{\mathcal{X}_i\}_{i=1}^{M}\}_{j=1}^{N}$ are the messages transmitted from the jth agent (each with M features) to the ego agent. The network learns BEV representation from the camera observations via BEV feature encoding which is detailed below and outputs object detection results. Note that when $N = 1$, there is no collaboration and $\xi_{\Phi}(1)$ reflects the single-agent perception performance.

C. Details

BEV feature encoding. We adopt the BEV encoder backbone in BEVFormer [7] to obtain BEV representations and perform the 3D object detection task. BEVFormer has demonstrated state-of-the-art performance in single-car 3D object detection tasks. This architecture is based on a Transformer-based design that incorporates efficient deformable attention layers as an alternative to traditional multi-head attention mechanisms. This choice mitigates the computational cost associated with modeling extensive sequences, such as pixels or voxels. In the deformable attention layer, 3D BEV queries are transformed into 2D reference points and projected onto the respective camera's features. Therefore, it is notable that when applied to multi-agent collaboration, communication overhead scales with the number of 3D-to-2D queries, increasing in tandem with the number of partner vehicles and cameras involved. To achieve efficient collaboration, we augment deformable attention module with a coordinate-based active selection network to assign each query with an interest score $\mathcal{I} = \{\mathcal{I}_{ki}\}_{k=1,i=1}^{N_{\text{car}},N_{\text{view}}}$ where $\mathcal{I}_{ki} \in [0,1]$. Here N_{car} and N_{view} mean the total number of agents and image views for each agent. We use Attn to denote each query point's attention weights result following Eq. 1:

$$\text{Attn}\left(q, F_{ki}\right) = \text{DfmAttn}\left(q, \mathcal{P}_k(q,i,j), F_{ki}\right) \tag{5}$$

And based on that we develop the pose-guided selective attention (PSA) layer:

$$\text{PSA}\left(q, F\right) = \sum_{k=1}^{N_{\text{car}}} \frac{1}{|\mathcal{R}_{\text{hit}}|} \sum_{i\in\mathcal{R}_{\text{hit}}} \mathcal{I}_{ki} \sum_{j=1}^{N_{\text{ref}}} \text{Attn}\left(q, F_{ki}\right) \tag{6}$$

Here $q \in \mathbb{R}^{1\times C}$ directly follows Eq. 2. Note the difference is that, $\mathcal{P}_k(q,i,j)$ is specifying projection matrix of the i-th camera of k-th vehicle denoted as $\mathbf{T'_k} \cdot \mathbf{T_i}$ with image feature F_{ki} , indicating car-to-car then ego-to-image transform. $\mathcal{I}_{ki} = 0$ means that query q will not attend to this camera image. We introduce how $\mathcal{I}$ is obtained in the below.

Active selection. As mentioned above, the active selection mechanism outputs the interest score $\mathcal{I}_{ki} \in [0,1]$ via a simple network. As illustrated in Fig. 2, it takes as input the BEV query q, along with the aforementioned transformation matrices $\mathbf{T_{c_{ki}}} = \mathbf{T'_k} \cdot \mathbf{T_i}$. The network that outputs interest score map $\mathcal{I}$ is formulated as:

$$\mathcal{I} = \sigma\left(\text{MLP}(\text{concat}(q, \text{PE}(\mathbf{T_{c_{ki}}}))\right) \tag{7}$$

Here PE$(\cdot)$ embeds transform matrix to a pose embedding, and σ represents sigmoid non-linearity which serves as a gated selection module. The goal is to generate a collection of interest scores for each query q with regards to every feature F_{ki} of the i-th camera from k-th car then use it to guide collaboration. From another perspective, this also forms a BEV map $\mathcal{M}_{ki} \in \mathbb{R}^{H\times W}$ for each feature indicating what queries are interested in it, as visualized in Fig. 3 (A) in the experiment section. And naturally for query q at (x,y): $I_{ki} = \mathcal{M}_{ki}(x,y)$. This map serves to highlight the relevance and significance of each feature for the ego's perception. The sigmoid gating ensures the resulting interest scores will be lining towards 0 or 1. During inference, for those queries with low interest scores, we set a threshold $\epsilon \ll 1$ and only select those that have the value above it. More formally, the query should collaborate with the i-th image of k-th vehicle only if $\mathcal{I}_{ki} > \epsilon$.

Task specific decoder head. Under collaborative conditions, we generate collaborative BEV encoding with PSA. Subsequently, the BEV encoding is fed into a Deformable DETR head [36] for 3D bounding boxes prediction.

Further experience shows that the collaborative BEV encoding significantly improves 3D detection performance on the basis of collaboration, and effectively leverages information from different perspectives.

IV. Experiments

A. Experiment setup

Dataset. We conducted experiments using the V2X-Sim Dataset [8], an extensive dataset that simulates complex urban driving scenarios using the CARLA simulator [37]. Our training dataset comprises 80 scenes, while the validation and testing dataset consist of 10 scenes each. The dataset is sampled at a rate of 5 Hz. Furthermore, we adapted the V2X-Sim Dataset to match the data format of the nuScenes standard [38] in the MMDet3D framework [39]. We preprocessed the voxel grids within a range of $[-51.2m, 51.2m]$ in the x and y-axes and lifted $N_{\text{ref}} = 4$ points on the z-axis. We also aimed to test our method on real-world datasets. However, current options have some limitations. The DAIR-V2X dataset [40] focuses on vehicle-to-infra collaboration, which is not applicable to our case. Additionally, the vehicle-to-vehicle dataset V2V4Real [41] has not yet released the camera data. We look forward to testing our method on real-world data once it becomes available in the future.

Baseline. We conducted a direct comparison between ActFormer and the BEVFormer version 1 model, which was extended to collaboration using all queries. We refer

TABLE I

COLLABORATIVE 3D OBJECT DETECTION RESULTS WITH DIFFERENT NUMBERS OF VEHICLES USING NUSCENES CENTER DISTANCE METRIC.

Paradigm	Method	AP with different N_{car}				
		1	2	3	4	5
Single-agent	BEVFormer [7]	52.1	N/A	N/A	N/A	N/A
Multi-agent	Co-BEVFormer	**52.1**	53.8	55.0	56.1	60.8
	ActFormer	51.7	**54.8**	**55.8**	**58.9**	**61.2**

TABLE II

COLLABORATIVE 3D OBJECT DETECTION RESULTS FOR DIFFERENT NUMBERS OF VEHICLES EVALUATED WITH BOUNDING BOX AP@IOU.

Method	mAP	AP@IoU=0.5 / AP@IoU=0.7 with N_{car}				
		1	2	3	4	5
Co-BEVFormer	@0.5	**32.54**	34.15	38.74	43.34	45.88
	@0.7	**19.76**	22.16	24.03	26.05	29.89
ActFormer	@0.5	31.40	**37.42**	**40.41**	**44.70**	**53.33**
	@0.7	19.71	**25.48**	**31.23**	**36.31**	**45.15**

to this baseline as Co-BEVFormer. This comparison allows us to assess the impact of the proposed active selection mechanism. The results of this comparison under different numbers of cars are presented in Table II. We also compare the performance with some existing collaborative perception models. They rely on different collaboration strategies and are all based on LiDAR inputs. All methods are listed in Table III for a comprehensive comparison. It is important to note that our study represents a novel approach for efficient camera-only collaborative perception, which distinguishes it from previous efficiency methods based on other modalities.

Implementation details. Our implementation consists of a 3-layer Pose-Guided Selection Attention (PSA) network and a 3-layer temporal self-attention mechanism from the backbone of BEVFormer, serving as the encoder for our framework. For Pose Embedding (PE), we embed the 4x4 homogeneous matrix into a 256-dimensional vector, matching the feature embedding dimension. Each query feature is concatenated with the Pose Embedding of the corresponding image. Then the result of concatenation is fed to the active selection network that consists of two linear layers to produce an interest score for each query. During inference, we remove queries with interest scores less than a predefined threshold, chosen as $\epsilon = 0.01$. This selection operation effectively reduces the number of queries communicated among the vehicles, resulting in approximately a 40% reduction in computations and operations.

Evaluation metrics. For the 3D object detection task, we report the Average Precision (AP) under two different bounding box Intersections Over Union (IOU) thresholds, namely 0.5 and 0.7, following the settings of previous works. Additionally, an alternative evaluation metric is utilized in the nuScenes dataset, which uses center distance difference as the mAP threshold. As BEVFormer was originally tested under this setting, we also evaluated our method using the same metric for a fair comparison. Results are listed in table I. Our performances under the two types of metrics appear to be consistent.

TABLE III

ACTFORMER COMPARED WITH OTHER COLLABORATION METHODS ON V2X-SIM FOR THE DETECTION TASK. "L" AND "C" INDICATE LIDAR AND CAMERA, RESPECTIVELY.

Method	Modality	Detection	
		AP@IoU=0.5	AP@IoU=0.7
When2com [3]	L	44.02	39.89
Who2com [4]	L	44.02	39.89
Where2comm [5]	L	59.10	52.20
V2VNet [16]	L	68.35	62.83
DiscoNet [1]	L	**69.03**	**63.44**
Co-BEVFormer	C	45.88	29.89
ActFormer	C	**53.33**	**45.15**

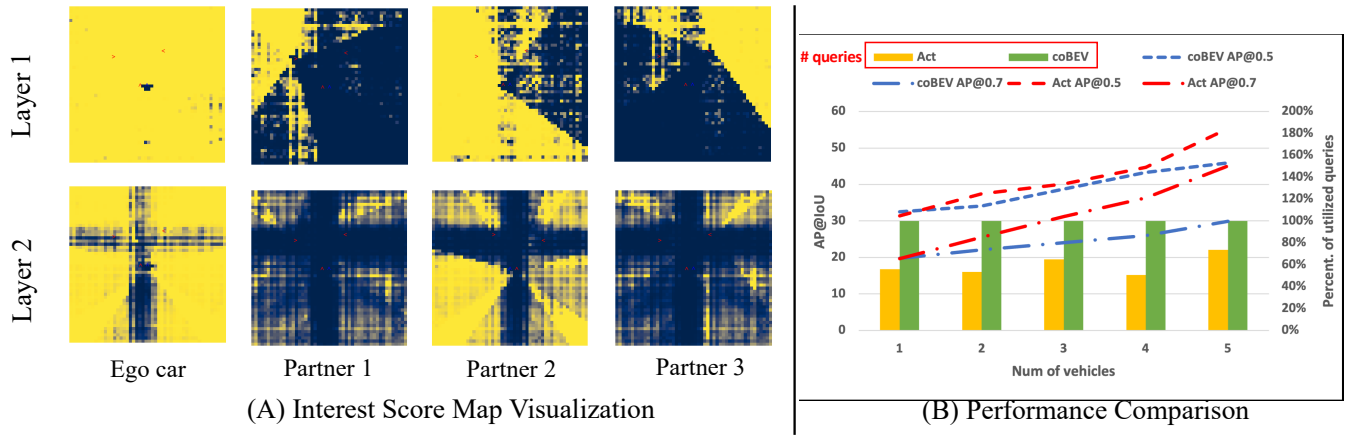
B. Quantitative results

Comparison with BEVFormer. We employed the BEVFormer as the single-agent detection model, without any collaboration, and refer to it as the *single-agent* BEVFormer. This model achieved an mAP of 52.1 under the nuScenes metric, which is reasonably competitive compared to the results reported in the original paper. For multi-vehicle, we trained the same backbone with inputs from all participating vehicles' images combined, which we term as *Co-BEVFormer*. The detection results for both models are presented in Tables I and II with different metrics. It is evident that ActFormer demonstrates performance improvements through collaboration and outperforms Co-BEVFormer, which simply aggregates multi-agent observations as input. The performance margin becomes even more pronounced as the number of vehicles increases, particularly under the challenging AP@IoU=0.7 metric.

Comparison with the state of the arts. Further, we conducted a comprehensive comparison of our camera-based approach with several strong LiDAR-based methods, including When2Com, Who2Com, DiscoNet, where2comm, and V2VNet, as presented in Table III. Despite the inherent challenges of performing 3D object detection from 2D images compared to LiDAR-based detection, ActFormer surpasses two LiDAR-based baselines and outperforms the camera-based baseline. This achievement can be attributed to the development of a spatial-information-guided interest score map for queries, facilitating the exchange of only pertinent messages and thereby ensuring efficient and effective collaboration. It is also important to note that this BEV query-based active collaboration approach sets our work apart from previous efforts in the field. Table IV provides an illustrative comparison of the different message fusion methods employed in these collaboration systems.

Efficiency.

The proposed query selection method not only leads to substantial performance improvements but also significantly reduces the number of query points, by approximately 50%, compared to the non-active utilization of all queries. This reduction not only minimizes communication overhead but also alleviates the computational burden, as the number of



(A) Interest Score Map Visualization (B) Performance Comparison

Fig. 3. (A) visualizations of the interest score map of ego vehicle and its 3 partners for 2 layers of PSA. (B) a comparison of the percentage of queries used versus the performance gain under AP@IoU evaluation. Act stands for ActFormer and coBEV stands for the baseline Co-BEVFormer.

TABLE IV

COMPARISONS OF COMMUNICATION APPROACHES IN COLLABORATIVE PERCEPTION SYSTEMS, FOLLOWING THE DEMONSTRATION IN [5].

Method	Message	Message fusion
When2com [3]	Full feature map	Attention per-agent
V2VNet [16]	Full feature map	Average per-agent
DiscoNet [1]	Full feature map	MLP attn per-location
Where2comm [5]	Sparse feature map	Attention per-location
ActFormer	Active BEV queries	Deformable attn per-query

TABLE V

COMPARISON ON EFFICIENCY. N_{ORI} AND N_{ACT} STAND FOR THE NUMBER OF NON-ACTIVE QUERIES OF ORIGINAL APPROACHES AND THAT OF ACTIVE QUERIES OF ACTFORMER, RESPECTIVELY.

	Components of different N_{car}				
	1	2	3	4	5
N_{ori}	4.45k	15.8k	23.5k	27.3k	31.5k
N_{act}	2.50k	8.46k	15.3k	13.7k	23.2k
$P_{N_{\text{car}}}$	55.98%	53.28%	65.01%	50.53%	73.60%

queries directly impacts the computation complexity in the subsequent attention mechanism. The percentage of query reduction for different numbers of agents compared to the Co-BEVFormer baseline is listed in Table V, presented as the percentage $P_{N_{\text{car}}} = N_{\text{act}}/N_{\text{ori}}$. Here, N_{ori} and N_{act} indicate the average number of the original non-active queries and the average number of active queries for all agents, respectively. These results are averaged over the testing dataset, demonstrating that our method reduces query points for all agents, even for a single agent.

C. Qualitative results

We present visualization results of the interest score maps in Fig. 3. Figure 3 (A) illustrates the total interest score maps at two active attention layers for both the ego car and its three partner vehicles. The interest score maps of all cameras on the same vehicles are combined together for visualization. The color map ranging from blue to yellow represents the interest score, ranging from 0 to 1, indicating the level of interest for the active queries. Notably, in the first layer, the ego car chooses queries to attend to each partner vehicle in areas that are farther away from the ego's own observation. Specifically, the blue regions on the interest score maps for partners overlap with the ego's viewpoints, which have already been covered by the ego's selections, as shown by the almost entirely yellow interest score map on the left. This behavior aligns with the logic of collaboration: by querying observations that are more challenging to access, the ego car gathers additional information. In the second layer, the network learns to identify crossroads, indicating that it has acquired a general understanding of the scene beyond the specific object detection task. It focuses its attention on vehicles within the intersection and objects outside the road. This observation underscores the significance of the interaction between BEV queries and the spatial information provided by each partner.

Figure 3 (B) provides a visual comparison of the percentages of queries utilized when different numbers of vehicles are collaborating, along with the corresponding detection performance. It is evident that as more vehicles participate, ActFormer consistently uses significantly fewer queries than the baseline and achieves a larger margin in detection performance. This clearly demonstrates the efficiency and effectiveness of ActFormer for multi-agent object detection.

V. CONCLUSION

This paper proposes ActFormer, an efficient and scalable method for multi-robot collaborative 3D object detection from 2D images with active 3D-to-2D queries. It utilizes the poses of collaborating partners and actively selects a sparse set of BEV queries to interact with the 2D image features for BEV representation learning. Comprehensive experiments prove that it significantly reduces information redundancy while still enhancing the detection performance. We believe ActFormer is a versatile method for multi-agent perception and plan to extend it to multi-modality input and different perception tasks in our future works.

References

[1] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021. 1, 2, 5, 6

[2] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *6th Annual Conference on Robot Learning*, 2022. 1, 2

[3] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115. 1, 2, 5, 6

[4] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883. 1, 2, 5

[5] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022. 1, 2, 5, 6

[6] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191. 1, 2

[7] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18. 1, 2, 4, 5

[8] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022. 2, 4

[9] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *European Conference on Computer Vision*. Springer, 2022, pp. 316–332. 2

[10] S. Su, Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao, "Uncertainty quantification of collaborative detection for self-driving," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5588–5594. 2

[11] Y. Li, Q. Fang, J. Bai, S. Chen, F. Juefei-Xu, and C. Feng, "Among us: Adversarially robust collaborative perception by consensus," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 186–195. 2

[12] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9243–9252. 2

[13] S. Su, S. Han, Y. Li, Z. Zhang, C. Feng, C. Ding, and F. Miao, "Collaborative multi-object tracking with conformal uncertainty propagation," *IEEE Robotics and Automation Letters*, 2024. 2

[14] Y. Li, Z. Lyu, M. Lu, C. Chen, M. Milford, and C. Feng, "Collaborative visual place recognition," *arXiv preprint arXiv:2310.05541*, 2023. 2

[15] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-robot collaborative perception with graph neural networks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2289–2296, 2022. 2

[16] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 605–621. 2, 5, 6

[17] Y. Li, J. Zhang, D. Ma, Y. Wang, and C. Feng, "Multi-robot scene completion: Towards task-agnostic collaborative perception," in *Conference on Robot Learning*. PMLR, 2023, pp. 2062–2072. 2

[18] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524. 2

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. 2

[20] C. Feichtenhofer, Y. Li, K. He, *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35 946–35 958, 2022. 2

[21] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 2

[22] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2147–2156. 2

[23] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *British Machine Vision Conference*, 2019. 2

[24] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082. 2

[25] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1521–1529. 2

[26] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3d object detection leveraging accurate proposals and shape reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 867–11 876. 2

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2

[28] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9087–9098. 2

[29] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proceedings of the European Conference on Computer Vision*, 2020. 2

[30] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564. 2

[31] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022. 2

[32] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[33] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern*

recognition, 2022, pp. 13 760–13 769. 2

[34] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803. 2

[35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773. 2

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. 2, 4

[37] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16. 4

[38] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631. 4

[39] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," https://github.com/open-mmlab/mmdetection3d, 2020. 4

[40] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370. 4

[41] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 712–13 722. 4