CenterCoop: Center-Based Feature Aggregation for Communication-Efficient Vehicle-Infrastructure Cooperative 3D Object Detection

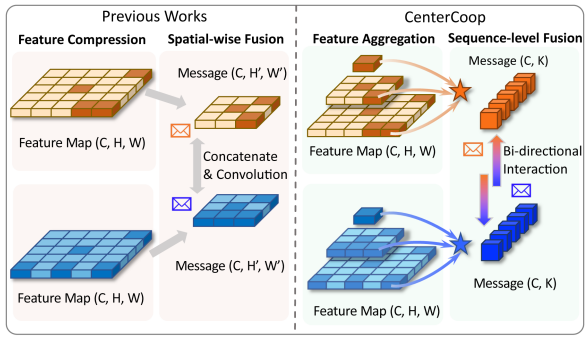
Linyi Zhou, Zhongxue Gan, and Jiayuan Fan

***Abstract*—Vehicle-Infrastructure Cooperative (VIC) 3D object detection is a challenging task for balancing communication bandwidth and detection performance. Intermediate fusion is recently studied to reach a better balance by transferring feature maps. Existing works mainly perform spatial-wise fusion and adopt feature compression to alleviate bandwidth cost by high-resolution feature maps, which would inevitably lead to information loss. Besides, overlapping observations between the two sensors would lead to near-duplicate detections, making trivial improvement to cooperative task while causing unnecessary bandwidth cost. To mitigate these problems, we propose a novel feature aggregation framework called CenterCoop, which first encodes the informative cues from the whole Bird's Eye View (BEV) context into compact center representations, enabling feature aggregation at sequence-level to significantly reduce the communication cost. Furthermore, to tackle the redundancy of transmitted data, we incorporate communication-aware regularization which enforces the network to extract complementary and beneficial cues for collaboration task. From an information-theoretic perspective, the proposed auxiliary constraints facilitate cooperative-view independence mining, resulting in enlarged perception range within the limited bandwidth. Extensive experiments on the DAIR-V2X dataset demonstrate the superior performance-bandwidth trade-off of CenterCoop, which achieves the state-of-the-art detection performance with less than 10% bandwidth cost.**

***Index Terms*—Computer vision for transportation, sensor fusion, object detection, segmentation and categorization.**

I. Introduction

RECENTLY, autonomous driving has unarguably been a trending topic which has drawn much attention from the community. With the rapid development of deep learning techniques, single-vehicle perception has achieved great improvements [1], [2], [3]. Nevertheless, its performance still suffers from limited perception range, occluded and sparse observations, decreasing the robustness and safeness of the perception system. To enhance the perception capability of each vehicle individual, roadside infrastructure emerges as a powerful collaborator [4], which provides a broader viewpoint with less occlusions, enlarging the perception range and covering more blind spots. However, incorporating infrastructure data is non-trivial due to specific problems such as bandwidth cost and temporal asynchrony. As a key challenge, reducing communication cost between individuals is desired for less time delay and potentially supports more agents to participate in collaboration. Thus, some works dive into the bandwidth-constrained problem to design bandwidth-efficient collaboration strategies [5], [6], [7].



Previous Works		CenterCoop
Spatial-wise	Message Fusion Type	Sequence-level
Low	Bandwidth Efficiency	High
✗	Multi-level clues	✓
✗	Fine-grained Interaction	✓

Fig. 1. Compared with previous fusion strategies for cooperative detection, the proposed CenterCoop performs finer-grained feature aggregation and interaction while achieving better bandwidth-efficiency. To avoid directly fusing high-resolution feature maps, previous works mainly adopt feature compression at first, which would inevitably lead to information loss. As a contrast, CenterCoop first globally aggregates multi-level cues from BEV grids into compact center embeddings, then performs message fusion at sequence-level.

To achieve a balance between performance and bandwidth, intermediate fusion is recently explored by [5], [6], [8] as a promising strategy. However, fusing intermediate features both efficiently and effectively remains a challenging problem. The reasons are twofolds. 1) High-resolution feature maps are required to preserve fine-grained cues for accurate prediction, resulting in huge bandwidth consumption. To reduce the bandwidth, existing works [5], [6], [7], [8] mainly perform feature compression before fusion. However, compression would inevitably lead to information loss in BEV representations (see Fig. 1). 2) Noisy factors already exist in single-view observations, and could aggravate spatial misalignment in cooperative-view fusion due

Manuscript received 30 July 2023; accepted 17 November 2023. Date of publication 5 December 2023; date of current version 5 March 2024. This letter was recommended for publication by Associate Editor and Editor upon evaluation of the reviewers' comments. This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0103, in part by the Shanghai Engineering Research Center of AI & Robotics, Fudan University, China, and in part by the Engineering Research Center of AI & Robotics, Ministry of Education, China. *(Corresponding authors: Zhongxue Gan; Jiayuan Fan.)*

The authors are with the Academy for Engineering and Technology, Fudan University, Shanghai 200437, China (e-mail: lyzhou21@m.fudan.edu.cn; ganzhongxue@fudan.edu.cn; jyfan@fudan.edu.cn).

Digital Object Identifier 10.1109/LRA.2023.3339399

to motion shift, sensor heterogeneity, projection errors, etc [6]. How to maintain the spatial correspondence while filtering sensor-related distractions has not been well investigated.

Apart from data fusion strategy, there exists another challenging factor which would cause unnecessary communication cost, i.e., the mutual redundancy between collaborators. As the perception range of vehicle and infrastructure sensors are usually partially overlapped, the extracted BEV features would share similar representations, resulting in near-duplicate predictions. Intuitively, this can be problematic under bandwidth-constrained scenarios. There are two underlying reasons. First, redundancy is not supposed to be transmitted where noisy cues may entangle with task-relevant features. Second, overlapping features may not contribute to performance gain when single-view features are already deterministic, thus causing unnecessary bandwidth cost. Recently, some work [8], [9] have verified that suppressing the redundancy is critical for good perception results. Nevertheless, how to exploit the mutual relationship to reduce communication cost remain untouched.

To tackle these challenges, we present CenterCoop, a novel feature aggregation framework for communication-efficient fusion, which consists of two key components. 1) Center-based Feature Aggregation (CFA) first performs single-view aggregation to initialize center embeddings and then conducts cooperative-view interaction to capture point-wise correlations between views. Specifically, for single-view aggregation, we exploit deformable attention to holistically accumulate spatial and long-range cues, which selectively samples multi-level BEV grids as attending keys, with less tendency to attend to noisy and task-irrelevant factors brought by each sensor. Subsequently, aggregated query sequences are fused by cooperative-view interaction, which utilizes bi-directional attention to enhance center representations. 2) Communication-aware regularization (CR) is further proposed to mitigate near-duplicate predictions caused by overlapping sensor observations. Mutual Information (MI) minimization is introduced to encourage mutual exclusiveness between views, assisted by another information-theoretic constraint to guarantee improved performance. These two terms jointly enable the model to achieve better detection performance with less communication cost.

In a nutshell, our contributions are three-fold:

- We propose a novel center-based feature aggregation framework for VIC 3D object detection task, which leverages center representations to inherently reduce the bandwidth. Specifically, it extracts compact yet informative center features through single-view aggregation and cooperative-view interaction.
- We propose a communication-aware regularization via mutual information minimization and an information gain constraint, which jointly facilitate complementary feature aggregation and guarantee the performance gain through collaboration. By encouraging the mutual exclusiveness of center heatmaps, communication cost is further reduced without performance degradation.
- The proposed CenterCoop achieves the state-of-the-art detection performance with less than 10% bandwidth cost on the real-world dataset DAIR-V2X, outperforming previous state-of-the-art works by a large margin.

II. Related Work

In this section, we first navigate LiDAR-based 3D detection methods designed for single-vehicle. Next, we compare fusion strategies for cooperative perception and discuss intermediate fusion which have a better performance-bandwidth balance. Finally, we introduce mutual information estimation as a potential to tackle data redundancy.

1) LiDAR-based 3D Object Detection: LiDAR-based 3D object detection methods can be mainly divided into point-based [10], [11], voxel-based [12], [13], point-voxel-based [14], [15] and projection-based [12], [16], [17]. PointRCNN [10] uses a bottom-up strategy to generate 3D proposals and learn semantic cues, which avoids using predefined boxes. PointPillars [12] converts voxels into pillars, and projects feature onto a BEV plane. [16] and [17] learn 3D representations from range image view. Recent works [2], [18] have shown the effectiveness of center-based 3D detectors on point cloud data. CenterPoint [2] detects objects using a center heatmap and performs bounding box regression using center feature representation. CenterFormer [18] is motivated by [2] and uses deformable attention to fuse multi-frame features. However, the success of these single-vehicle detectors has not been extended to collaborative detection task. We use center-based representations to design cooperative 3D detectors.

2) Intermediate Fusion: Fusion strategies for cooperative perception can be divided into input-based early fusion [19], [20], [21], output-based late fusion [22], [23] and feature-based intermediate fusion [5], [6], [7], [8]. As early fusion consumes large bandwidth for transmitting raw data and late fusion suffers from sharing noisy detection results, intermediate fusion emerges as a promising strategy to better balance the performance and bandwidth cost. V2X-ViT [6] uses transformer to fuse intermediate features, where compression is performed on the channel dimension using 1×1 convolution. To alleviate the large bandwidth cost, CRCNet [8] uses a convolutional autoencoder to compress feature maps into lower resolutions. Where2comm [5] proposes to use a spatial confidence map for selectively transmitting feature in high-confidence regions to reduce the huge communication cost by spatial-wise fusion. These methods mainly focus on designing compression strategies for spatial-wise feature fusion, while we leverage center-based representation to enable sequence-level feature fusion, which essentially mitigates the communication cost.

3) Mutual Information Estimation: Mutual Information (MI) is a measure of the relationship between two random variables that has drawn increasing attention in various research topics [8], [24], [25], [26], [27]. A majority of works exploit MI maximization by estimating its lower bounds to achieve enhanced complementary information [25], [26], [28]. [25] adopts mutual information maximization to align temporal features for video-based human pose estimation, which maximizes the task-relevant cues mined from supporting frames. MI minimization, which requires an upper bound, is also studied in various fields such as information bottleneck [29] and disentangled representation learning [24], [30]. [26] uses MI minimization to reduce the information redundancy for alleviating the artifacts in Pan-sharpening. Zhang et al. [27] introduce mutual

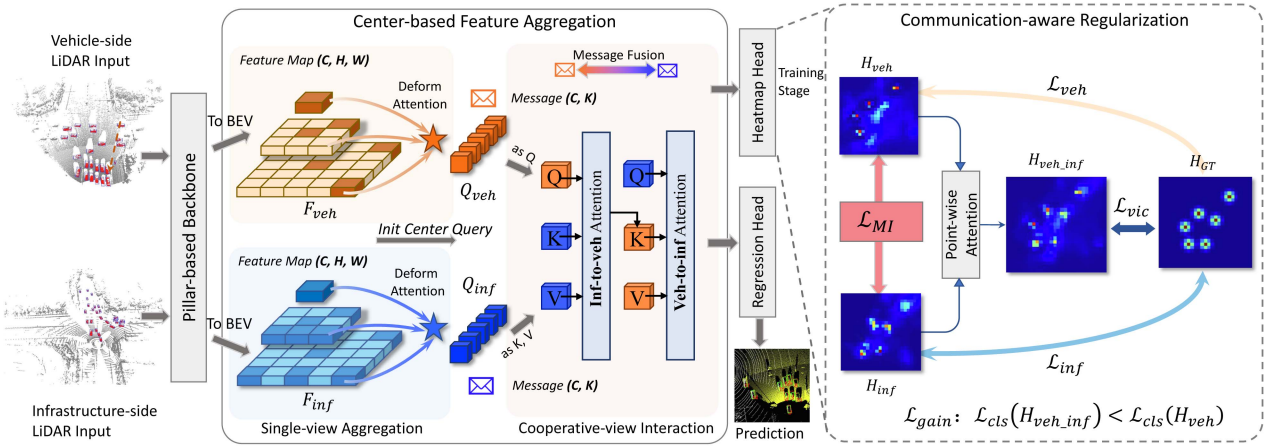


Fig. 2. Overview of the proposed CenterCoop. It achieves communication-efficient VIC 3D object detection via Center-based Feature Aggregation (CFA) and Communication Regularization (CR). Firstly, paired LiDAR inputs from infrastructure-side and vehicle-side are processed into high-dimensional BEV feature maps F_{veh} and F_{inf} of shape (C, H, W). Secondly, CFA performs single-view aggregation on multi-level BEV grids to initialize compact center queries Q_{veh} and Q_{inf} of shape (C, K). These queries are exchanged as messages followed by fusing cooperative-view interaction to achieve bi-directional feature enhancement. In the training stage, the proposed CenterCoop engages CR to further reduce the bandwidth cost, where the heatmap supervision is jointly regularized by mutual information minimization $\mathcal{L}_{MI}$ to suppress the cooperative-view redundancy and the information gain constraint $\mathcal{L}_{\text{gain}}$ to enforce complementary aggregation for beneficial collaboration.

information minimization to facilitate the multi-modal feature fusion by enhancing the complementary cues between RGB image and depth data. CRCNet [8] utilizes mutual information minimization in the cooperative perception task to suppress the redundancy and noisy factors between agents. Nevertheless, how to use mutual information to reduce the communication cost for efficient cooperative perception has not been investigated.

III. Method

In order to achieve better performance-bandwidth trade-off, we present CenterCoop (as shown in Fig. 2), which consists of two main components: Center-based Feature Aggregation (CFA) and Communication-aware Regularization (CR). Given LiDAR frames input I_{veh} and I_{inf}, CFA exploits center queries to facilitate feature extraction, transmission and interaction under bandwidth limit. CR works in the training stage to encourage complementary distribution of center heatmap proposals, contributing to less bandwidth consumption. Note that in the inference stage, only center features are needed for bounding box regression thus heatmaps are not transferred. CFA builds on CenterFormer [18], which provides details on the DETR-based pipeline and center-based detection. We tailors CenterFormer to work on cooperative task with two distinctive aspects. 1) CFA performs sequence-level fusion on cross-view features while CenterFormer performs spatial-wise fusion on multi-frame features. 2) CFA integrates bi-directional feature interaction to better capture fine-grained relationships between views.

A. Center-Based Feature Aggregation Network

CenterPoint [2] is a pioneering work which detects objects as centers from point cloud data, without the handcrafted anchor design. Although center representations have been adopted, the predictions are generated from 2D feature map, making it impractical to fuse sequence-level features. Existing methods use FPN-like architectures to fuse spatial features [5], which will, however, lead to extra communication cost. CenterFormer [18] makes it practical for directly predicting boxes from center embeddings, however, its data fusion scheme is still spatial-wise. As designed for multi-frame detection, features of previous timestamps are concatenated then fused by 3×3 convolution. Thus 2D feature maps of supporting frames still need to be transmitted for message fusion. To make it practical for sequence-level fusion, a straight-forward solution is to directly concatenate center embeddings for prediction. However, this will lead to suboptimal results for lacking the interaction between views.

The proposed CenterCoop uses single-view aggregated center embeddings to initialize compact yet informative queries for cooperative-view interaction. Spatial correspondence is encoded into center embeddings in the first stage, then sequence-level fusion is performed to capture global pair-wise dependences between views.

1) Single-view Feature Aggregation: Given BEV feature maps $F_{\text{veh}} \in \mathbb{R}^{C \times H \times W}$ and $F_{\text{inf}} \in \mathbb{R}^{C \times H \times W}$ obtained from the pillar-based feature extractor [12], we aim to extract finer-grained features for each side with less noisy factors included. As mentioned in Section I, high-resolution feature maps are required to preserve fine-grained cues for accurate detections. Besides, long-range and multi-level cues are also critical for sparse observations near the overlapped perception boundaries. Those far and sparse objects are desirable to be improved by the supporting view. Inspired by the recent proposed deformable attention [31] which can naturally aggregate multi-level features, we leverage this sparse attention to perform single-view feature aggregation. The benefits are: 1) View-specific distractions, such as sensor heterogeneity, projection errors and motion shift can be alleviated by adaptively sampling multi-level BEV grid locations as attending keys. 2) Long-range and multi-level cues are gathered from whole BEV context, contributing to compact yet informative center queries for feature fusion.

Similar to CenterFormer [18], we take top-K centers from the predicted heatmap as the initialized center queries $Q_{\text{veh}} \in \mathbb{R}^{C\times K}$ and $Q_{\text{inf}} \in \mathbb{R}^{C\times K}$ for each side, attending to keys sampled from multi-level BEV feature grids. Here C represents feature channels and K refers to the number of queries. Given a reference center location p at all heads and feature levels, deformable cross attention uses a linear layer to obtain 2D sampling offsets Δp. The feature at $p+\Delta p$ will be extracted as the attending key through bilinear sampling. Positional encodings are obtained by a linear layer to encode center locations. Formally, the output embeddings are enhanced through multi-level deformable attention for each single-view:

$$\text{DeformAttn}(p) = \sum_{m=1}^{M} W_m \left[\sum_{l=1}^{L} \sum_{k=1}^{K} \sigma\left(W_{mlk}\mathcal{C}(p)\right) x^l\left(p+\Delta p_{mlk}\right) \right] \quad (1)$$

where m indexes the attention head, l indexes the input feature level, and k indexes the sampling point. x^l represents the multi-level BEV feature at level l, $\mathcal{C}(p)$ is the center feature, and $\sigma(W_{mlk}\mathcal{C}(p))$ is the attention weight.

2) Cooperative-view Feature Interaction: After single-view aggregation which extracts informative features from the whole BEV context, we obtain compact query sequences from infrastructure-side Q_{inf} and vehicle-side Q_{veh} of shape (C, K) along with their positions P_{veh} and P_{inf} of shape $(1, K)$, representing the top-K BEV locations of highest-confidence score. These compact queries and positions are packed as messages for the cooperative-view interaction module. Since the single-view enhanced features contain less noisy cues, it is beneficial to perform pair-wise interaction to capture complementary features of both sides. Motivated by attention mechanism in [32], where each element of the query sequence attends to all elements of the key-value sequence, we adopt this global attention to aggregate complementary information from both sides.

As is shown in Fig. 3, CenterCoop performs bi-directional interaction to progressively enhance cooperative-view representations. In contrast to single-vehicle detector DETR [1], where queries are initialized with learnable parameters and attending keys are taken from the whole feature map, CenterCoop makes two adjustments for cooperative detection. First, bi-directional interaction is performed via inf-to-veh cross-attention and veh-to-inf cross-attention. Second, for each cross-attention module, queries are initialized with center features from one side (e.g., Q_{veh}) and attending keys are comprised of center features from the other side (e.g., Q_{inf}).

To get mutual augmented representations, CenterCoop first performs infrastructure-to-vehicle (inf-to-veh) attention, then the intensified center queries Q_{veh} are used as the key for vehicle-to-infrastructure (veh-to-inf) attention. Formally, we use Q_i to denote the set of query sequence, K_i and V_i for key and value sequence, where i and j are the indices. For inf-to-veh cross-attention, i indexes the K-dimensional query feature Q_{veh}, j indexes the K-dimensional source feature Q_{inf}, and vice-versa.

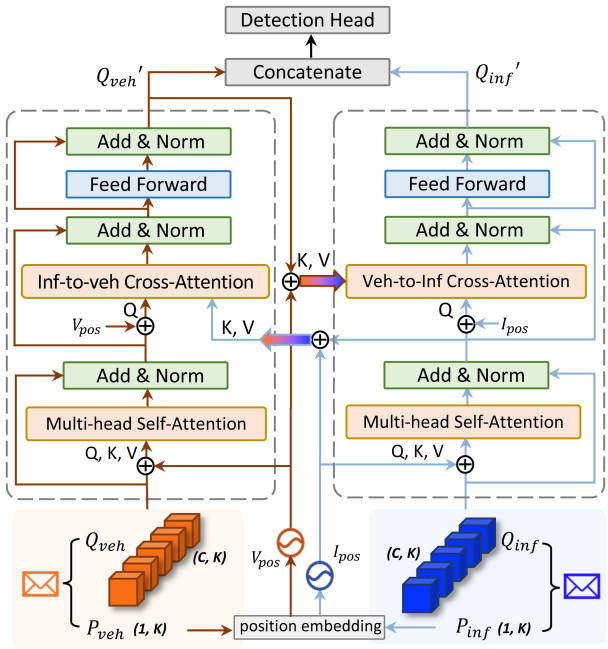


Fig. 3. Detailed network structure of the proposed cooperative-view interaction module. Given messages of center queries and their positions, it performs bi-directional feature fusion to capture complementary cues via inf-to-veh cross-attention and veh-to-inf cross-attention.

Let f^q and f^k be the query and key feature. For inf-to-veh attention, $f^q = Q_{\text{veh}}$ and $f^k = Q_{\text{inf}}$, or vice versa. We adopt a linear layer to encode P_{veh} and P_{inf} into position embeddings, denoted as V_{pos} and I_{pos} respectively. Formally, $V_{pos} = \text{Linear}(P_{\text{veh}})$ and $I_{pos} = \text{Linear}(P_{\text{inf}})$. To refer to specific elements within embeddings, we introduce the notations E_i^{pos} and E_j^{pos}. Here, E_i^{pos} denotes the i^{th} element of V_{pos}, and E_j^{pos} denotes the j^{th} element of I_{pos}. The multi-head attention is formulated as:

$$\text{Attn}(Q,K,V) = \sum_{m=1}^{M} W_m \left[\sum_{j\in\Omega_j} \sigma\left(\frac{Q_i K_j}{\sqrt{d}}\right) \cdot V_j \right] \quad (2)$$

$$Q_i = f_i^q W_q^m + E_i^{pos}, K_j = f_j^k W_k^m + E_j^{pos}, V_j = f_j^k W_v^m \quad (3)$$

where m is the head index, σ is the softmax function, d is the feature dimension, and W is the learnable weight. Ω_j is the the set of key elements, comprised of center features from the supporting view. Finally, augmented Q'_{veh} and Q'_{inf} are concatenated for bounding box regression. Note that in our experiment, $K \ll HW$ thus the global attention would not lead to heavy computational cost.

B. Communication-Aware Regularization.

As the perception range of collaborators are usually partially overlapped, single-view enhanced queries Q_{veh} and Q_{inf} may contain geometrically nearby features. This would lead to redundant detections and cause unnecessary communication cost. Hence, CenterCoop further incorporate auxiliary constraints to facilitate complementary feature learning.

1) Mutual Information Minimization: Mutual information is widely studied [24], [25], [27] to measure the nonlinear statistical independence. Different from previous works that use MI to maximize the complementary cues, we use MI minimization to explicitly encourage the cooperative-view independence, i.e., learning features that are positionally mutual exclusive. As a cooperative learning task, each collaborator should provide supplementary cues rather than focusing on salient but dupilcated objects that are already covered by the other view. Ideally, center heatmap serves as a probabilitic distribution reference to measure the geometrical complementarity of two BEV representations. Given the predicted heatmap H_{inf} and H_{veh}, the Mutual Information (MI) is formulated as:

$$MI\left(H_{\text{veh}}, H_{\text{inf}}\right) = S\left(H_{\text{veh}}\right) + S\left(H_{\text{inf}}\right) - S\left(H_{\text{veh}}, H_{\text{inf}}\right) \quad (4)$$

where $S(.)$ represents the Shannon entropy, $S(H_{\text{veh}})$ and $S(H_{\text{inf}})$ are marginal entropies, and $S(H_{\text{veh}}, H_{\text{inf}})$ is the joint entropy of H_{veh} and H_{inf}. Intuitively, we have the Kullback-Leibler divergence (KL) of the two latent variable (or the conditional entropies) as:

$$KL\left(H_{\text{veh}} \| H_{\text{inf}}\right) = S_{H_{\text{inf}}}\left(H_{\text{veh}}\right) - S\left(H_{\text{veh}}\right), \quad (5)$$

$$KL\left(H_{\text{inf}} \| H_{\text{veh}}\right) = S_{H_{\text{veh}}}\left(H_{\text{inf}}\right) - S\left(H_{\text{inf}}\right), \quad (6)$$

where $S_{H_{\text{inf}}}(H_{\text{veh}}) = -\sum_x H_{\text{veh}}(x) \log H_{\text{inf}}(x)$ is the cross-entropy. Summing over (4), (5) and (6), we obtain:

$$\begin{aligned} MI\left(H_{\text{veh}}, H_{\text{inf}}\right) &= S_{H_{\text{inf}}}\left(H_{\text{veh}}\right) + S_{H_{\text{veh}}}\left(H_{\text{inf}}\right) \\ &- S\left(H_{\text{veh}}, H_{\text{inf}}\right) - \left(KL\left(H_{\text{veh}} \| H_{\text{inf}}\right) + KL\left(H_{\text{inf}} \| H_{\text{veh}}\right)\right). \end{aligned} \quad (7)$$

Given the data from infrastructure and vehicle-side, $H(h_{\text{veh}}, h_{\text{inf}})$ is nonnegative, then MI objective is given by:

$$\begin{aligned} \mathcal{L}_{MI} &= \left(S_{H_{\text{inf}}}\left(H_{\text{veh}}\right) + S_{H_{\text{veh}}}\left(H_{\text{inf}}\right)\right) \\ &- \left(KL\left(H_{\text{veh}} \| H_{\text{inf}}\right) + KL\left(H_{\text{inf}} \| H_{\text{veh}}\right)\right). \end{aligned} \quad (8)$$

In the context of vehicle-infrastructure collaboration, $MI(H_{\text{veh}}, H_{\text{inf}})$ reflects the dependence between vehicle-side and infrastructure-side data (H_{inf}). Intuitively, $MI(H_{\text{veh}}, H_{\text{inf}})$ measures the reduction of uncertainty in vehicle-side predictions when infrastructure-side data is observed, or vice versa. By minimizing MI, we explicitly encourage the positional complementarity to generate proposals from different regions. Since overlapped centers is reduced to make the most of the query number K (which represents the allocated bandwidth), the perception range can be effectively enlarged within limited bandwidth.

2) Information Gain Constraint: Although MI minimization can encourage the mutual exclusiveness of cooperative-view heatmaps, extracted features may not guarantee the improved detections by incorporating the infrastructure data. Inspired by the minimal sufficient statistics and the Information Bottleneck theory in [33], we further integrate an information gain constraint to keep task-relevant information sufficient when reducing the mutual information.

Given single-view heatmap H_{veh} and H_{inf}, we fuse them to get the cooperative heatmap $H_{\text{veh\_inf}}$ using dot product self-attention. Denote the cooperative ground-truth heatmap as H_{GT}, classification loss $\mathcal{L}_{cls}(\cdot)$ is measured between the predicted heatmap and H_{GT}. We enforce the cooperative loss $\mathcal{L}_{cls}(H_{\text{veh\_inf}})$ is smaller than vehicle loss $\mathcal{L}_{cls}(H_{\text{veh}})$ by a threshold δ_{thr}. Formally, the information gain constraint is:

$$\mathcal{L}_{\text{gain}} = \mathcal{L}_{cls}\left(H_{\text{veh}}\right) - \mathcal{L}_{cls}\left(H_{\text{veh\_inf}}\right) - \delta_{thr} \quad (9)$$

Given vehicle's data, $\mathcal{L}_{\text{gain}}$ enforces a performance gain can be achieved by adding infrastructure-side data. After feature fusion, the resulting heatmap $H_{\text{veh\_inf}}$ should lead to decreasing loss, indicating higher detection performance. With a controllable number of query K, CenterCoop simultaneously reduces the bandwidth cost and improves the detection performance via mining the cooperative-view independence.

3) Overall Optimization: Focal loss [34] and L1 Loss are utilized for heatmap classification and box regression, denoted as $\mathcal{L}_{cls}(\cdot)$ and $\mathcal{L}_{reg}$ respectively. Single-view heatmap losses $\mathcal{L}_{cls}(H_{\text{veh}})$ and $\mathcal{L}_{cls}(H_{\text{inf}})$ are also incorporated, serving as a guidance for learning task-relevant cues to enhance perception of each side. The total heatmap loss is:

$$\mathcal{L}_{hm} = \mathcal{L}_{cls}(H_{\text{veh\_inf}}) + w_{\text{veh}}\mathcal{L}_{cls}(H_{\text{veh}}) + w_{\text{inf}}\mathcal{L}_{cls}(H_{\text{inf}}) \quad (10)$$

where w_{veh} and w_{inf} are balancing factors for single-view losses. Finally, the overall optimization is formulated as:

$$\mathcal{L} = \lambda_{hm}\mathcal{L}_{hm} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{MI}\mathcal{L}_{MI} + \lambda_{\text{gain}}\mathcal{L}_{\text{gain}} \quad (11)$$

where λ_{hm}, λ_{reg}, λ_{MI} and λ_{gain} are the trade-off factors for each term.

IV. Experiments

A. Setup

1) Dataset: DAIR-V2X [4] is the first real-world dataset for VIC 3D cooperative detection task, with all frames captured from real scenarios with 3D cooperative annotations. In this task, ego-vehicle receives and fuses information from infrastructure to enhance the detection result around itself. Infrastructure side data is collected by 300-beam LiDAR and vehicle-side data is collected by 40-beam LiDAR. DAIR-V2X provides VIC-Sync and VIC-Async dataset for cooperative task. We study the synchronous scenario and use the VIC-Sync dataset, where the timestamps between frame pairs are less than 30 ms. VIC-Sync is split into train/val/test part as 5:2:3. Previous works [4], [5] report results on the publicly-available validation set. For fair comparison, we use the same training set (4811 pairs) and validation set (1789 pairs).

2) Evaluation metric: VIC 3D object detection task is formulated as integrating infrastructure-side information to enhance the perception capability of each vehicle individual via message communication. The task has two primary goals: 1) better detection performance, i.e., higher Average Precision (AP); 2) less communication bandwidth, i.e., lower Average Byte (AB). The official evaluation metric is AP with BEV IoU threshold 0.5 for cars. We follow [4] to evaluate the communication AB.

TABLE I
COMPARISON WITH THE SOTA METHODS ON DAIR-V2X DATASET

Model	Fusion	AP(IoU=0.5)↑ 3D	AP(IoU=0.5)↑ BEV	AB(Byte) ↓
DAIR-V2X [4]	Late	41.90	47.96	336.16
DAIR-V2X [4]	Early	50.03	53.73	1382275.75
DAIR-V2X [4] †	Late	56.06	62.06	478.61
DAIR-V2X [4] †	Early	62.61	68.91	1382275.75
Where2comm [5]	Mid	58.46	-	598088.73
Where2comm [5]	Mid	63.54	-	3091766.00
Where2comm [5]	Mid	63.71	-	12892243.44
CenterCoop (K=30)	Mid	62.18	72.28	**62160**
CenterCoop (K=100)	Mid	**64.72**	**75.24**	207200

† denotes the updated results reported on the official site. Mid means intermediate fusion.

3) Implementation Details: The perception range is [0 m, 99.84 m] for the X axis, [−40.32 m, 40.32 m] for the Y axis, and [−3 m, 1 m] for the Z axis. "x" direction is only front since [4] does not label objects outside the camera's view. In the training stage, objectives in (10) are balanced through trade-off factors w_{veh} and w_{inf}, which are both set to 0.5 empirically. For the balance factor of regularization, λ_{MI} and λ_{gain} is set to 0.3 and 0.2 respectively. δ_{thr} in (9) is set to 0.5. We have not heavily tuned these parameters. The feature channel of queries C is set to 256. For detailed architecture, we get multi-scale BEV features using PointPillars backbone [12] with a grid size of (0.24 m, 0.24 m). Single-view Aggregation consists of one transformer layer using deformable cross-attention setting heads $M = 3$ in (1). Cooperative-view interaction consists of two transformer layer using vanilla cross-attention setting heads $M = 4$ in (2). Center-based detection head is the same as [18]. We train all models using NVIDIA GeForce RTX 3090 GPU.

B. Quantitative Results

To validate the effectiveness of the proposed CenterCoop, we conduct experiments on the real-world DAIR-V2X dataset (VIC-Sync) and the results are shown in Table I.

In Where2comm [5], metric AB is not reported, we calculate the corresponding value following the [5] and [4], which is given by $volume = \log(AB/2)$. Apart from center features and their corresponding indices (as mentioned in Section III-A), messages of CenterCoop also include class labels and scores for bounding box regression, both are K-dimensional vectors of shape $(1, K)$. Thus the total communication AB is $(C + 3) * K * 8$. Note that queries only need to be transferred from infrastructure to the vehicle since the task is defined to enhance the ego-vehicle perception [4].

As compared with other methods, CenterCoop achieves the best detection results as well as the communication AB, demonstrating its superior performance-bandwidth trade-off. In terms of detection results, CenterCoop surpasses all previous methods on both 3D and BEV AP (64.72% and 75.24% respectively). As for communication AB, CenterCoop only costs less than 1/10 bandwidth compared to [5]. When allocating smaller bandwidth (598088.73), the detection performance of Where2comm significantly drops to 58.46%. In contrast, our model maintains good detection performance with a smaller decline (75.24% → 72.28%), while saving 70% bandwidth (207200 → 62160).

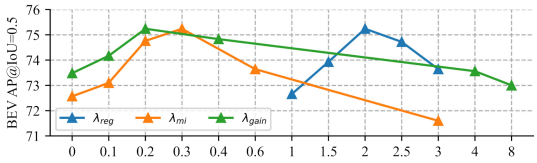


Fig. 4. Sensitivity analysis on λ_{reg}, λ_{MI} and λ_{gain}. (K=100).

TABLE II
COMPARISON WITH DIFFERENT INTERMEDIATE FUSION STRATEGIES

Model	AP(IoU=0.5)↑ 3D	AP(IoU=0.5)↑ BEV	AB(Byte) ↓
CenterPoint-FPN	60.94	69.23	23482368
CenterFormer-VehOnly	54.10	63.59	0
CenterFormer-MTF-3×3	**64.74**	74.14	17891328
CenterFormer-MTF-1×1	50.36	63.76	17891328
CenterCoop (w/o CFI)	55.11	65.84	207200
CenterCoop (K=100)	64.72	**75.24**	207200
CenterCoop (K=50)	63.13	73.42	103600
CenterCoop (K=30)	62.18	72.28	**62160**

We also conduct sensitivity analysis for trade-off terms in (11), setting λ_{hm} to 1 and K to 100. Results in Fig. 4 indicate that the performance of CenterCoop is not sensitive to these hyper-parameters within a reasonable range.

C. Ablation Studies.

1) Effect of Fusion Strategy: To demonstrate the effectiveness of CFA, we implement several center-based detection methods with different fusion strategies. 1) CenterPoint-FPN: use [2] and adopts dot-product self-attention fusion in feature pyramid network. 2) CenterFormer-VehOnly: use single-frame CenterFormer [18] with only the vehicle-side data. 3) CenterFormer-MTF-3 × 3: use multi-frame CenterFormer [18], where multi-frame spatial features are concatenated then fused by 3×3 convolution, transferring HW attending keys for feature fusion. 4) CenterFormer-MTF-1 × 1: the same configuration as 2) except using 1 × 1 convolution. 5) CenterCoop (w/o CFI): we replace the cooperative-view feature interaction (CFI) module with direct concatenation of Q_{veh} and Q_{inf} for box regression. 6) CenterCoop: the proposed model with both CFA and CR.

Results are shown in Table II. CenterPoint-FPN achieves a moderate performance through directly fusing features via grid-wise attention. The multi-level fusion has the largest AB cost. CenterFormer-MTF-3 × 3 has a high detection performance, indicating the effectiveness of deformable attention. However, AB is still large due to the spatial-wise fusion scheme. Note that when we replace the 3 × 3 convolution with 1 × 1, the performance has a significant drop, revealing that the spatial misalignment exists in cooperative-view and larger perception field is necessary for effective fusion. By virtue of single-view aggregation, the misalignment can be naturally alleviated since multi-level features are gathered to initialize the center queries for effective fusion. Without fine-grained feature interaction, CenterCoop (w/o CFI) has lower performance than CenterPoint-FPN,

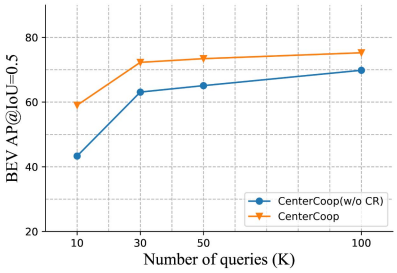


Fig. 5. Performance-bandwidth trade-off comparison for the ablation on CR module. Bandwidth cost is proportional to K.

TABLE III
ABLATION ON THE EFFECTIVENESS OF $\mathcal{L}_{MI}$ AND $\mathcal{L}_{\text{gain}}$

	$\mathcal{L}_{MI}$	$\mathcal{L}_{gain}$	AP@3D (IoU=0.5)	AP@BEV (IoU=0.5)
(a)			59.28	69.83
(b)	√		62.20	73.48
(c)		√	59.87	72.57
(d)	√	√	**64.72**	**75.24**

validating the effectivess of CFI. CenterCoop achieves the best performance-bandwidth trade-off with sequence-level fusion. For $K = 100$, it achieves best 75.24% BEV AP within 207200 AB. For $K = 30$, it achieves 72.28% BEV AP within only 62160 AB. It achieves comparable detection performance (64.72% 3D AP) compared to CenterFormer-MTF-3×3 (64.74% 3D AP), while consuming much smaller bandwidth.

2) Effect of Communication-aware Regularization: To validate the effectiveness of CR, we evaluate the detection performance variance of within different bandwidth limit, setting K to 10, 30, 50 and 100 respectively. Results are shown in Fig. 5. It indicates that CR helps CenterCoop achieve better performance-bandwidth trade-off. We further decompose CR to verify the effectiveness of each constraint, setting K = 100. Results are shown in Table III. Several conclusions can be drawn. First, when MI is incorporated, the detection performance is significantly boosted (3D AP +2.92%, BEV AP +3.65%). This can be ascribed to the promoted mutual exclusiveness and suppressed redundancy for effective collaboration. When integrated with $\mathcal{L}_{\text{gain}}$, the model performance is also improved by guaranteeing the cooperative-view information gain. Jointly optimizing these constraints yields significantly better detection performance (3D AP 59.28%→64.72%, +5.44% and BEV AP 69.83%→75.24%, +5.41%), revealing the complementary effect of two losses. Intuitively, mutual information minimization serves as a guidance for mining complementary cues for improved detections.

D. Qualitative Results

1) Detection Visualization: We visualize our results in Fig. 6. Sparsity and occlusion of point cloud can be observed from each view, as well as misalignment caused by view-specific distractions. Note that label noise also exists in VIC-Sync dataset. With the proposed strategy, CenterCoop can achieve accurate detections on these challenging cases (see Fig. 6(a)).

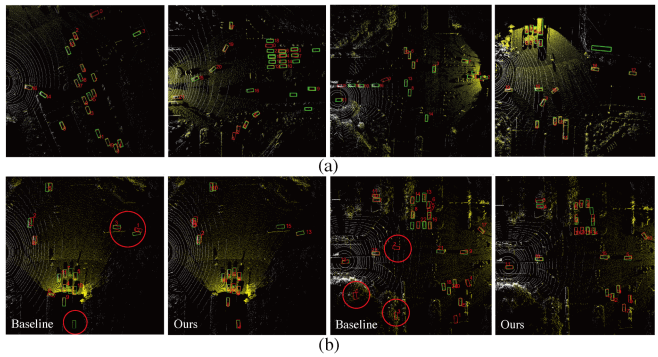


Fig. 6. Detection results on DAIR-V2X. White and yellow points represent LiDAR input from vehicle and infrastructure. Ground truth and predicted bounding boxes are marked as green and red. (a) shows results on challenging cases and (b) compares with the DAIR-V2X baseline.

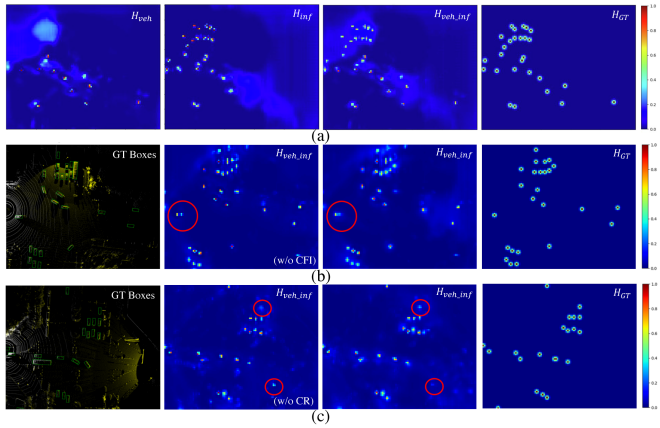


Fig. 7. Heatmap visualization of CenterCoop model. (a): Effective collaboration by leveraging complementarity between two views; (b): Ablation on Cooperative-view Feature Interaction (CFI) and (c): Ablation on Communication-aware Regularization (CR).

Moreover, compared with DAIR-V2X baseline [4], CenterCoop predicts more complete and accurate detections for sparse and occluded objects, while generating less noisy predictions (see Fig. 6(b)).

2) Heatmap Visualization: We further analyze the intermediate heatmap predictions. Results are shown in Fig. 7. When comparing H_{veh}, H_{inf} and $H_{\text{veh\_inf}}$, we observe that complementary cues of each view are effectively captured to achieve more complete detections (see Fig. 7(a)) with enlarged perception range. Moreover, we analyze the challenging case where fast-moving objects lead to center shift (see Fig. 7(b)). Directly concatenating single-view features (w/o CFI) causes near-duplicated detections, while CFI captures this relation and achieves more robust result in $H_{\text{veh\_inf}}$ through bidirectional enhancement. We further examine the effect of CR. Results in Fig. 7(c) show that when incorporating CR, 1) more complementary cues are captured and less redundancy is observed; 2) uncertainties are suppressed and less noisy predictions are observed. This validates the effectiveness of CR which aims to suppress mutual redundancy and encourage beneficial aggregation.

V. Conclusion

VIC 3D object detection is challenging for balancing perception performance and bandwidth. We propose CenterCoop to tackle this problem by transmitting compact center features as queries instead of the whole feature map. The proposed CenterCoop achieves the SOTA detection result and saves more than 90% bandwidth. Specifically, it consists of 1) center-based feature aggregation and 2) communication-aware regularization. Center-based representations are first leveraged to tackle the restricted bandwidth limit, through single-view feature aggregation and cooperative-view feature interaction. Moreover, we explicitly encourage the mutual exclusiveness as well as collaborative task-relevant cues through jointly optimizing the two communication-aware regularization terms, i.e., mutual information minimization and the information gain constraint.

Acknowledgment

The authors would like to thank Wenchao Ding and Ke Wu for the invaluable ideas, discussions, and visualizations that contributed to this letter.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, and Kirillov, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[2] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.

[3] Z. Zhou et al., "SGM3D: Stereo guided monocular 3D object detection," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10478–10485, Oct. 2022.

[4] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21329–21338.

[5] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *in Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 4874–4886.

[6] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. 17th Eur. Conf.*, 2022, pp. 107–124.

[7] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29541–29552.

[8] G. Luo, H. Zhang, Q. Yuan, and J. Li, "Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 3578–3586.

[9] J. Florez-Lozano and M. Gongora, "Cooperative and distributed decision-making in a multi-agent perception system for improvised land mines detection," *Inf. Fusion*, vol. 64, pp. 32–49, 2020.

[10] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.

[11] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in *Proc. CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11037–11045.

[12] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12697–12705.

[13] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4490–4499.

[14] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "STD: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1951–1960.

[15] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10529–10538.

[16] Y. Chai, P. Sun, J. Ngiam, and W. Wang, "To the point: Efficient 3D object detection in the range image with graph convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16000–16009.

[17] Z. Liang, Z. Zhang, M. Zhang, X. Zhao, and S. Pu, "RangeioUDet: Range image based real-time 3D object detector optimized by intersection over union," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7140–7149.

[18] Z. Zhou, X. Zhao, Y. Wang, and P. Wang, "CenterFormer: Center-based transformer for 3D object detection," in *Proc. 17th Eur. Conf.*, 2022, pp. 496–513.

[19] E. Arnold,, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.

[20] S. Aoki, T. Higuchi, and O. Altintas, "Cooperative perception with deep reinforcement learning for connected vehicles," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 328–334.

[21] G. Luo, H. Zhang, H. He, J. Li, and F.-Y. Wang, "Multiagent adversarial collaborative learning via mean-field theory," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 4994–5007, Oct. 2021.

[22] G. Volk, A. v. Bernuth, and O. Bringmann, "Environment-aware development of robust vision-based cooperative perception systems," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 126–133.

[23] H. Liu, P. Ren, S. Jain, M. Murad, M. Gruteser, and F. Bai, "Fusioneye: Perception sharing for connected vehicles and its bandwidth-accuracy trade-offs," in *Proc. IEEE 16th Annu. Int. Conf. Sensing, Commun., Netw.*, 2019, pp. 1–9.

[24] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[25] Z. Liu et al., "Temporal feature alignment and mutual information maximization for video-based human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10996–11006.

[26] M. Zhou, K. Yan, J. Huang, Z. Yang, X. Fu, and F. Zhao, "Mutual information-driven pan-sharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1788–1798.

[27] J. Zhang et al., "RGB-D saliency detection via cascaded mutual information minimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4338–4347.

[28] A. Sanghi, "Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 626–642.

[29] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1xwcyHFDr

[30] X. Hou, Y. Li, and S. Wang, "Disentangled representation for age-invariant face recognition: A mutual information minimization perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3692–3701.

[31] X. Zhu, W. Su, and L. Lu, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke

[32] A. Vaswani, N. Shazeer, N. Parmar, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[33] Y. Tian, C. Sun, B. Poole, D. Krishnan, and P. Isola, "What makes for good views for contrastive learning?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6827–6839.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.