ImVoxelENet: Image to Voxels Epipolar Transformer for Multi-View RGB-based 3D Object Detection

Gang Xu, Haoyu Liu, Biao Leng*, Zhang Xiong
Beihang University
* Corresponding author: lengbiao@buaa.edu.cn

Abstract

Three-dimensional object detection based on purely image data is a significant challenge in the field of computer vision. The core issue lies in how to accurately perform epipolar geometry matching across multiple views to obtain latant geometric priors. Existing methods establish correspondences along epipolar line features in voxel space through multiple layers of convolution. However, this step is often arranged in the later stages of the network, which limits the overall performance. To address this challenge, we introduce a novel framework that integrates geometric epipolar constraint, which we refer to as ImVoxelENet. We start from the back-projection of pixel-wise features and design an attention mechanism that captures the relationship between forward and backward features along the ray across multiple views. This approach enables the early establishment of geometric correspondences and structural connections between epiploar lines. On the multi-view detection dataset ScanNetV2, extensive comparative and ablation experiments demonstrate that our proposed network achieves a 1.1% improvement in mAP, highlighting its effectiveness in enhancing 3D object detection performance.

Keywords: 3D Object Detection, Multi-view Geometry, Transformer, Attention, Deep Learning

1. Introduction

3D object detection plays a crucial role in indoor scene understanding and is fundamental for a wide range of applications [23, 53, 1], such as robotics, augmented reality, and autonomous systems. Accurate 3D detection enables machines to perceive and interact with their environment more effectively, which is essential for tasks like navigation, manipulation, and scene reconstruction. For instance, in robotics [53], precise 3D detection allows for safe and efficient movement within dynamic environments. As indoor environments [24] are typically dense with objects and diverse in structure, a robust 3D detection system offers the potential to revolutionize how technology interfaces with real-world scenarios. While recent advances have shown promise in this field, especially with the integration of deep learning and multi-view approaches, achieving reliable and efficient 3D detection remains an area of active research. To address these issues, there is a growing need for innovative approaches that can efficiently fuse multi-view data and enhance 3D object detection performance in indoor environments [20, 46].

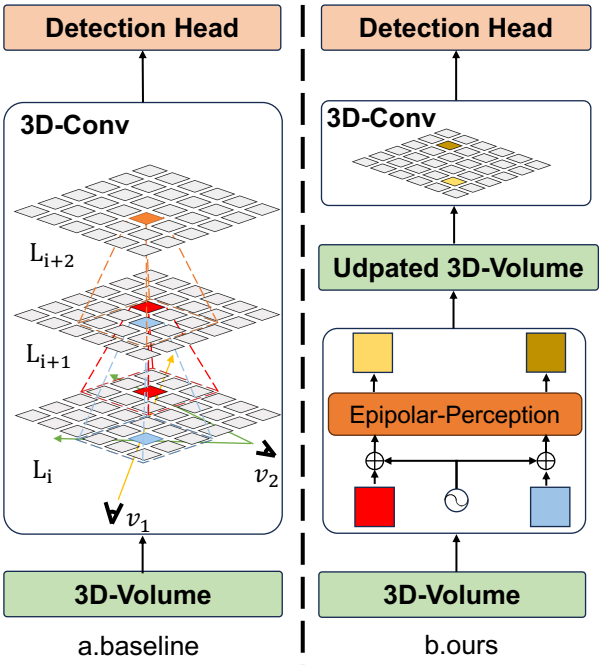

Figure 1. Difference on utilization of epipolar geometry in multi-view detection. In previous methods, as shown on the left, the relationships between distant voxels on the same epipolar line rely on multiple convolutions in voxel space to expand the receptive field. In contrast, as depicted in the right image, our method establishes the earlier front-to-back relationships along the same ray, followed by convolutions. The red and periwinkle colors respectively map the feature along a single epipolar line. For simplification, we represent the 3D convolution block by a 2D top-down view.

In indoor scenes [11, 33, 56, 45, 16, 50], multi-view 3D

object detection has emerged as a promising approach to accurately infer the 3D structure of objects by leveraging multiple 2D images captured from different viewpoints. The general strategy [22, 57] involves projecting 2D features from these images into a shared 3D space, allowing for the reconstruction of object shapes, positions, and orientations. This process typically includes stages such as feature extraction, multi-view aggregation, and 3D bounding box prediction. By integrating information from multiple perspectives, multi-view 3D detection can potentially overcome limitations faced by single-view methods [7, 47], such as occlusion, scale variation, and ambiguities in depth perception. Despite its potential, multi-view 3D object detection presents several challenges. One of the primary difficulties is the accurate alignment and fusion of multi-view information due to inconsistencies in appearance caused by variations in lighting [2], occlusion [25], and texture [14]. Indoor environments are particularly challenging because they often feature densely cluttered scenes with complex object arrangements [3, 49] and overlapping structures, making it harder to establish reliable correspondences between different views. Another challenge is the inherent noise and variability introduced by factors such as camera calibration errors and occlusions, which can affect the precision of the back-projection process when mapping 2D image features to a 3D space. A critical technical aspect of addressing these challenges is developing effective methods for handling the epipolar geometry between views to ensure accurate matching and fusion. Techniques such as epipolar line sampling and attention-based feature aggregation have shown promise in capturing long-range dependencies and establishing geometric correspondences across multiple views [41]. Additionally, learning-based methods that incorporate geometric priors, attention mechanisms, or ray-based sampling approaches have been developed to enhance the network's ability to infer 3D spatial relationships from 2D data. However, designing architectures capable of robustly capturing these multi-view relationships while maintaining computational efficiency remains an open research problem. As such, the development of more advanced methods for feature alignment, aggregation, and 3D spatial understanding is crucial for pushing the boundaries of multi-view 3D object detection in indoor settings.

Current multi-view 3D object detection methods exhibit limitations in handling epipolar matching effectively [29]. Specifically, they rely on multi-layer 3D convolutions to capture the relationships along the same ray, as shown in (a) of Fig. 1, which often occurs at later stages of the network. This delayed integration means that the network cannot establish the crucial geometric correspondences early on, resulting in suboptimal performance in capturing fine-grained spatial details. As a consequence, the ability to accurately interpret the front-to-back connections along each ray is hindered, which reduces the network's overall efficiency and effectiveness in achieving precise 3D object detection. Addressing this shortcoming is key to enhancing the performance of multi-view 3D detection in complex indoor environments.

Through our analysis, we identify that establishing a more precise matching logic for the epipolar relationship is the key challenge in this task. To enhance this relationship, we propose the ImVoxelENet method. Traditional approaches rely on late-stage 3D convolution modules with larger receptive fields to achieve global feature fusion, followed by classification and localization of features within voxels, as shown in (a) of Fig. 1. Inspired by the transformer architecture, we posit that leveraging an attention mechanism to capture the geometric relationships along the ray can facilitate earlier realization of spatial relationships compared to existing methods, as illustrated in (b) of Fig. 1. By establishing epipolar matching relationships through this mechanism, the network can more effectively harness geometric cues, enabling implicit spatial awareness and enhancing the understanding of spatial features throughout the network.

Specifically, we design the ImVoxelENet structure based on the ImVoxelNet [31] framework. First, we introduce an indexing mechanism for the rays within the original back-projection operation. We map the correspondence between each pixel in the multi-view image sequence and the voxel space. We then record the indices of these feature rays and select a few intersecting voxels along each ray. Next, we proceed to construct the complete 3D voxel feature space. All pixels from the entire image sequence are projected into the voxel space according to the camera's intrinsic and extrinsic parameters. The intersecting points of different rays within the voxels are fused using a feature-level numerical averaging approach. Subsequently, we develop a epipolar-based transformer structure. Using the previously established indices, we input the nearest and farthest voxel points along each ray to obtain updated voxel features. These updated features are then added as residuals to the original 3D voxel space. Finally, we apply subsequent 3D convolutional neck layers and detection heads to produce the final 3D object detection results.

Our contributions can be summarized in the following three aspects:

- We design a novel network framework, ImVoxelENet , for multi-view 3D object detection using only image data. This approach does not require any 3D geometric priors or supervision, yet it achieves significant performance improvements.

- We introduce a new ray-based transformer structure that implements a long-range attention mechanism along the ray before applying 3D convolutions. This

enhances the network's ability to perceive spatial structures, improving its overall understanding of the 3D environment.

- We conduct extensive experiments, including comparative studies with existing methods and ablation studies on our approach, which convincingly demonstrate the effectiveness and superiority of the proposed method.

2. Related Work

2.1. Point-based 3d object detection

Point clouds are inherently three-dimensional, making it seem intuitive to use a 3D convolutional network for detection. Zhang et al. [52] propose Fully Sparse TRansformer (FSTR) for efficient LiDAR-based 3D object detection, combining state-of-the-art sparse backbones and dynamic queries, achieving superior performance on nuScenes and Argoverse2 benchmarks. Chen et al. [5] introduce FocalFormer3D, utilizing Hard Instance Probing (HIP) to reduce false negatives in 3D detection, improving recall and outperforming benchmarks in detection and tracking tasks across LiDAR and multi-modal datasets. Real-Aug [13] presents a synthesis-based LiDAR augmentation method addressing unrealistic scan patterns in existing approaches, achieving state-of-the-art performance on nuScenes by prioritizing realistic LiDAR data generation and employing a real-synthesis training strategy.

However, these approaches demand extensive computational resources, leading to slow inference times, particularly for large outdoor scenes. Recent outdoor detection methods [44, 15] mitigate this issue by projecting the 3D point cloud onto the bird's-eye view (BEV) plane, significantly reducing runtime. A common technique in point cloud processing involves subdividing the point cloud into voxels. The projection onto the BEV plane implies that all voxels within a vertical column are encoded into a fixed-length feature map. This resulting pseudo-image can then be processed by a 2D object detection network to generate final predictions.

For indoor object detection, methods typically generate object proposals for each point in the point cloud. However, since many indoor objects are not convex, the geometric center of an object may not lie within the object itself (e.g., the center of a table or chair may be between its legs). As a result, proposals based solely on a single center point can be inaccurate. To address this, indoor detection methods rely on deep Hough voting to generate more reliable proposals [27, 26, 55].

2.2. RGB-based 3d object detection

To address the challenges of 3D object detection in outdoor environments, particularly in BEV (Bird's Eye View) autonomous driving scenarios, the computer vision community has developed a wide range of strategies and methods for effective object detection.

Far3D[12] propose a sparse query-based framework for efficient long-range 3D object detection from surround-view images, using 2D priors, perspective-aware aggregation, and range-modulated denoising, achieving state-of-the-art results on the Argoverse 2 dataset. The paper[18] introduce Ray Denoising, a plug-and-play module for multi-view 3D object detection that improves depth estimation accuracy by sampling along camera rays to create hard negative examples. It achieves a 1.9% mAP gain over the state-of-the-art on the NuScenes dataset, demonstrating strong generalization capabilities. [58] introduces Historical Object Prediction (HoP) for multi-view 3D detection, leveraging temporal information by generating pseudo Bird's-Eye View (BEV) features from historical timestamps. HoP improves BEV feature learning and achieves 68.5% NDS and 62.4% mAP on nuScenes, outperforming existing 3D detectors. Wang et al. present StreamPETR[38], a long-sequence modeling framework for multi-view 3D object detection. Utilizing an object-centric temporal mechanism, it achieves 67.6% NDS and 65.3% AMOTA on nuScenes, comparable to LiDAR-based methods. The lightweight version outperforms SOLOFusion by 2.3% mAP and is 1.8× faster in FPS. VCD[17] is a framework enhancing camera-only 3D object detection by leveraging an apprentice-friendly multi-modal expert model (VCD-E) and a fine-grained distillation module. VCD-A achieves state-of-the-art performance on nuScenes with a 63.1% NDS score. SparseBEV[19] presents a fully sparse 3D object detector that outperforms dense counterparts by enhancing adaptability in BEV and image space. It achieves state-of-the-art performance with 67.5 NDS on nuScenes and maintains real-time inference at 23.5 FPS. SparseAD[51] takes up with a sparse query-centric paradigm for end-to-end autonomous driving. It handles detection, tracking, and online mapping without dense BEV representation, achieving state-of-the-art full-task performance on the nuScenes dataset and reducing the gap with single-task methods. PolarBEVDet[48] is a multi-view 3D object detector using polar BEV representation instead of Cartesian BEV, tailored with a polar-view transformer, temporal fusion module, and detection head. It achieves superior performance on the nuScenes dataset, effectively handling view symmetry and image distribution. In VideoBEV[8], Han et al. propose a long-term recurrent fusion strategy for camera-based BEV 3D perception. It effectively combines rich long-term information with an efficient fusion pipeline, achieving strong results on nuScenes with 55.4% mAP in object detection, outperforming existing methods. [43] introduces CAPE, a novel method using Camera view Position Embedding for multi-view 3D object detection. CAPE operates under local camera-view coor-

dinates, enhancing 3D detection. It achieves state-of-the-art performance (61.0% NDS, 52.5% mAP) on nuScenes among LiDAR-free methods.

For indoor scenarios, ImVoxelNet [31] stands out as one of the most representative and effective methods for 3D object detection. ImVoxelNet is a prominent method for 3D object detection in indoor scenes. It utilizes multi-view images to project 2D features into a 3D voxel space, enabling the network to effectively capture spatial relationships. By constructing a 3D voxel grid from 2D image features, ImVoxelNet can perform efficient 3D convolutions, leading to accurate 3D object localization and detection. This voxel-based approach [39] allows the model to represent the 3D structure of indoor environments effectively, making it a strong performer in complex indoor detection tasks. Total3DUnderstanding[24, 11], takes a comprehensive approach to indoor scene understanding by simultaneously tackling 3D object detection, layout estimation, and shape reconstruction. Zhang et al. [50] propose an image-based local structured implicit network to enhance object shape estimation. Additionally, they refine the 3D object pose and scene layout through a novel implicit scene graph neural network, which leverages implicit local object features. Stekovic et al. [32] performs joint selection and optimization of proposals from a generated pool, aiming to minimize the objective term. In the initial application involving floor plan reconstruction from point clouds, this approach selects and refines room proposals represented as 2D polygons, optimizing based on an objective function that combines fitness as predicted by a deep network with regularization terms on room shapes. Wang et al. [35] introduce a novel augmented reality (AR) solution tailored for tele-meeting applications. This approach integrates neural networks with simultaneous localization and mapping (SLAM) techniques to attain comprehensive scene understanding and user localization solely from RGB images.

3. Motivation

Inputs and goals. Our method takes as input a collection of images $I_n \in \mathbb{R}^{W \times H \times 3}$, each with intrinsic and extrinsic camera parameters and arbitrary resolution, where t represents the n-th image in a set of N images. In this paper, we focus on multi-view scenarios where $T > 1$. The objective is to predict multiple 3D object bounding boxes using our proposed ImVoxelENet network framework. Each bounding box is parameterized as $(x, y, z, w, h, l,)$, where (x, y, z) represents the spatial coordinates of the center of the 3D bounding box, and w, h, l denote its width, height, and length, respectively.

Advantage of Multi-view 3D Detection We conducted an extensive review of existing image-based 3D object detection methods within scene understanding [22, 31, 12, 24, 21] and compared them with several prominent point cloud-based approaches [48, 15, 44], leading to the formulation of the research motivation for this paper.

Specifically, point cloud-based methods inherently incorporate explicit 3D spatial information in their input data. This type of data allows these methods to bypass the process of reconstructing the 3D spatial structure from images or other sources, which also helps avoid errors that might arise during such reconstruction. As a result, these methods tend to yield more accurate and reliable outcomes. Some approaches, although not utilizing point cloud data as input, explicitly reconstruct the geometric relationships between multiple views. They then perform detection based on the reconstructed point cloud. These methods often use 3D point cloud data as a supervisory constraint for the network, facilitating the learning of geometric shapes.

In contrast, the method proposed in our research directly uses 2D RGB images as input and conducts 3D object detection within the 2D image space. While this design increases the complexity of the network's learning process, it significantly reduces the constraints on the input data, as it does not rely on point cloud data. This not only makes the method more flexible in terms of data requirements but also allows for easier expansion to larger-scale training datasets. Consequently, our approach aims to strike a balance between performance and data accessibility, offering a practical solution for 3D object detection in scenarios where point cloud data is unavailable or difficult to obtain.

In our proposal, although introducing additional challenges in terms of learning and inference, has the potential to scale more effectively with larger datasets, providing a promising direction for future research and development in the field of 3D object detection.

Post-Perception by Convolution. To further improve the learning performance of our network, we explore related methods that utilize only RGB images. Through this investigation, image-based approaches commonly leverage epipolar geometry as implicit geometric cues, as illustrated in (a) of Fig. 2. Specifically, these methods typically map each image into a predefined 3D voxel grid using the intrinsic and extrinsic parameters of the camera [22, 31]. However, a frequent challenge with single-view images is the absence of depth information[42, 34]. As a result, multiple possible intersection points along the same ray in the voxel grid are often assigned identical features from a single pixel.

This situation leads to redundancy, as several voxels along the same ray may correspond to the same feature, which introduces the need for additional validation and the elimination of erroneous points. This objective is usually achieved by analyzing the correspondences from other viewpoints. In essence, multiple images are used to establish correspondence for specific points across different per-

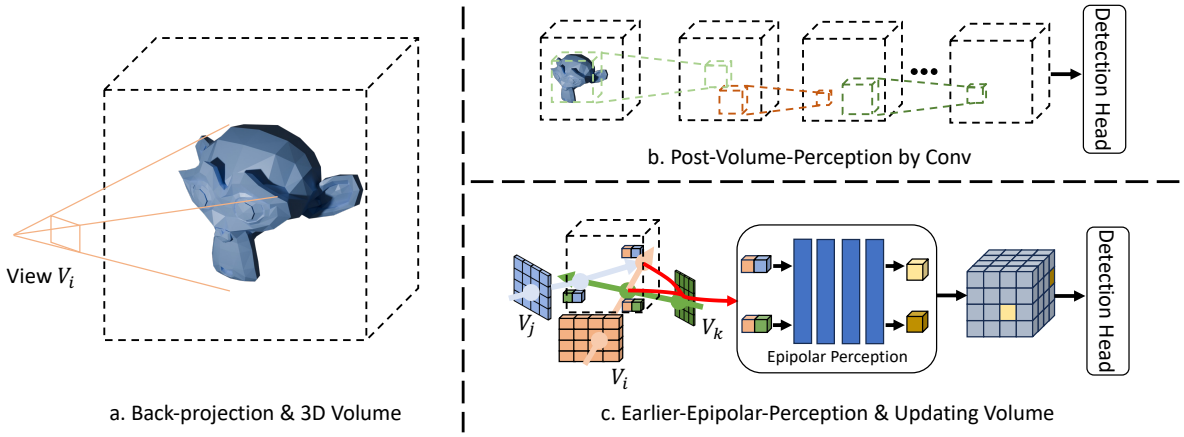

Figure 2. Illustration of Our Motivation: (a)The back-projection and construction of 3D feature voxels are general operations in researchings [22, 31], visualizing three possible matching epipolar lines and point features from three different viewpoints. (b)In existing approaches, multiple convolution layers are required to connect the features along the epipolar line within a larger receptive field. In contrast, our method (c), establishes these connections earlier through a specific architecture.

spectives, which helps to disambiguate the 3D location of features along the ray.

However, this matching process typically involves long-range relationships, as the corresponding points along a ray may be distributed at different depths. As shown in (b) in Fig.2 and Fig. 1, these methods often apply convolution operations after projecting the images into 3D space, but the effectiveness of this approach is constrained by the receptive field of the convolutional kernels. Since the receptive field grows incrementally with the network's depth, the model can only capture global correspondences at higher layers of the network. This limitation inherently restricts the ability to match long-range correspondences along the ray early in the process, thus impacting the overall efficiency and accuracy of the object detection process.

By addressing these challenges and understanding the limitations imposed by convolutional operations in these image-based methods, our research aims to propose improvements that can enhance the ability of the network to capture and utilize these long-range correspondences more effectively, potentially leading to more accurate 3D object detection from RGB images.

Earlier Perception by Transformer. To improve this process, we are supposed to develop strategies to address the long-range correspondences along the ray in both forward and backward directions [37, 54]. The logic behind this approach is straightforward: we propose an additional residual path alongside the conventional convolution logic used in neighboring spatial regions. As attention mechanisms have achieved remarkable success across various domains, significantly boosting performance, which is well-suited for situations requiring the establishment of feature relationships along rays with indeterminate lengths and voxel positions in 3D space. By incorporating attention mechanisms, we aim to enhance the learning capacity of the network by an earlier perception as illustrated in (c) of Fig. 2.

Specifically, we designed an attention-based module that utilizes a cross-attention mechanism to match and learn features associated with multiple voxel locations corresponding to the same pixel across different images. Importantly, rather than relying on the features of a single pixel's ray from a single image, we leverage the voxel features obtained from the projection of multiple images. This choice stems from the observation that feature correspondences derived solely from the front-to-back relationships in a single image are often meaningless; they do not reveal significant feature variations and offer no basis for selecting the correct spatial positions. The cross-attention mechanism can better discern feature differences and establish meaningful correspondences between voxels across different perspectives. This approach enhances the model's ability to filter out erroneous points and focus on relevant spatial relationships, ultimately improving the accuracy and robustness of the 3D object detection process. Furthermore, the residual path helps maintain gradient flow during training, allowing the model to effectively capture both local and global feature interactions without being constrained by the limited receptive field of standard convolutions.

In summary, this novel attention-based approach is designed to address the inherent challenges of long-range feature matching in 3D voxel space, providing a more powerful and flexible method for learning geometric relationships from multiple images. We believe this integration of attention mechanisms will significantly enhance the overall performance of our 3D object detection system.

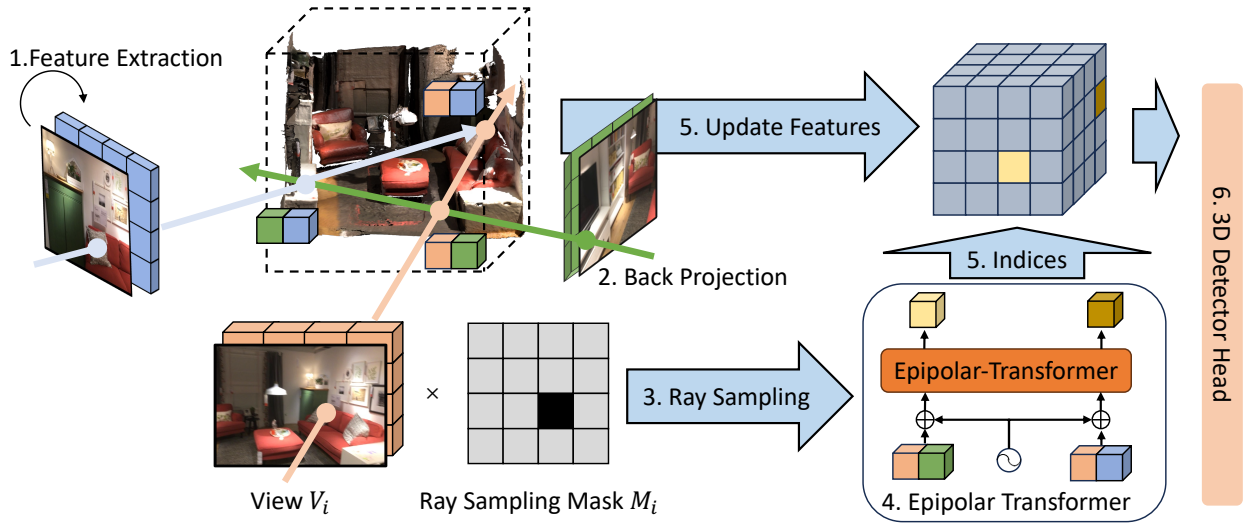

Figure 3. Pipeline of our ImVoxelENet , beginning by a 2D feature extractor from the input images, which are then projected into a 3D voxel space. Following this, we introduce a novel ray sampling and ray-transformer to capture the features along the ray. Then we update the original voxel space with the refined features. Finally, we apply a detection head to perform the 3D object detection task.

4. Method

In this section, we provide a detailed introduction to the proposed ImVoxelENet network model and methodology. The workflow of our network is illustrated in Fig. 3. In the following sections, we will formalize and define the task inputs, outputs, and specific scenarios by mathematical formulations in Section. 4.1. Starting from feature extraction, we propose a back-projection method oriented towards ray sampling, which will be elaborated on in Section 4.2. Then, in Section. 4.3, we will explain how ray-level sampling and indexing are performed and how attention mechanisms are used to establish long-range spatial-ray connections in Section. 4.4. Finally we index and update the original voxel features, leading to the final 3D object detection with corresponding losses to train our ImVoxelENet .

4.1. Back-projection

We follow [31, 22] to set our back-projection steps.

Let $I_t \in \mathbb{R}^{W \times H \times 3}$ represents the t-th image from a sequence of T images. In this literature, we focus on the situation containing more than 5 images to detect 3d objects in each scene. In accordance with the methodology proposed by Murez et al. [22], we initiate our process by extracting two-dimensional (2D) features from each input image by a pre-trained 2D backbone. This extraction yields four distinct feature maps, which are characterized by their dimensions: $\frac{W}{4} \times \frac{H}{4} \times c_0$, $\frac{W}{8} \times \frac{H}{8} \times 2c_0$, $\frac{W}{16} \times \frac{H}{16} \times 4c_0$, and $\frac{W}{32} \times \frac{H}{32} \times 8c_0$. These feature maps are subsequently aggregated through the utilization of a Feature Pyramid Network (FPN), resulting in a unified tensor F_t with dimensions $\frac{W}{4} \times \frac{H}{4} \times c_1$. It is important to note that c_0 and c_1 are parameters specific to the backbone network, with their exact values detailed in the implementation section.

The 2D feature representations F_t for the t-th input are then transformed into a three-dimensional (3D) voxel space, denoted as $V_t \in \mathbb{R}^{N_x \times N_y \times N_z \times c_1}$. The orientation of this 3D volume is defined such that the z-axis is oriented perpendicularly to the ground plane, while the x-axis points forward, and the y-axis is orthogonal to both the x- and z-axes. We empirically estimate the spatial boundaries along all three axes as $x_{\min}, x_{\max}, y_{\min}, y_{\max}, z_{\min}, z_{\max}$ followed [55, 15, 22]. With a predetermined voxel size s, the relationship between the number of voxels and the spatial extent can be expressed as $N_x s = x_{\max} - x_{\min}$, $N_y s = y_{\max} - y_{\min}$, and $N_z s = z_{\max} - z_{\min}$.

A pinhole camera model serves as the basis for establishing the correspondence between 2D coordinates (u, v) within the feature map F_t and 3D coordinates (x, y, z) within the voxel volume V_t:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \Pi \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & 1 \end{bmatrix} K R_t \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix},$$

where K and R_t denote the intrinsic and extrinsic matrices, respectively, and Π represents the perspective projection. Following the projection of 2D features into 3D space, all voxels lying along a given camera ray inherit the same set of features. Additionally, a binary mask M_t, sharing the same dimensions as V_t, is constructed to indicate whether a voxel lies within the frustum of the camera. For an individ-

ual image I_t, the mask M_t is defined as:

$$M_t(x,y,z) = \begin{cases} 1, & \text{if } 0 \le u < \frac{W}{4} \text{ and } 0 \le v < \frac{H}{4} \\ 0, & \text{otherwise.} \end{cases}$$

4.2. 3D Volume Construction

Subsequently, the 2D feature map F_t is projected onto the 3D voxel grid V_t for all valid voxels, according to:

$$V_t(x,y,z) = \begin{cases} F_t(u,v), & \text{if } M_t(x,y,z) = 1 \\ 0, & \text{otherwise.} \end{cases}$$

The final step involves aggregating the binary masks $M_1, \ldots, M_t$ to form a composite mask M, computed as:

$$M(x,y,z) = \begin{cases} \sum_t M_t(x,y,z), & \text{if } \sum_t M_t(x,y,z) > 0 \\ 1, & \text{otherwise.} \end{cases}$$

Ultimately, the 3D voxel volume V is synthesized by averaging the projected features across the valid voxels in the individual volumes $V_1, \ldots, V_t$, following the formula:

$$V = \frac{1}{M} \sum_t M_t V_t.$$

4.3. Ray Sampling

The above steps in last section follow the standard procedures established by ImVoxelNet [31] and AtlasNet [22], maintaining a nearly identical structural framework. Next, we implement an independent ray sampling mechanism, which is divided into two components, as illustrated in the Fig. 4. The first component involves random selection based on pixel weights for each image, while the second component focuses on front-and-back sampling along the projected rays. Below, we will elaborate on the rationale behind this design, detailing the mechanisms involved and outlining the logical framework for the practical implementation.

In the first module, our design objective is to sample the number of rays. Since the purpose of extracting rays and their features is to utilize them in the subsequent Epipolar-Transformer stage, it is necessary to impose a certain limit on their quantity. This limitation helps conserve GPU memory, allowing the network to train more stably. To achieve this, we first utilize the effective mask mentioned in the previous chapter to perform element-wise multiplication with the image and its features, identifying points in a single photo that can be mapped to the voxel space. We then extract all valid pixels from all images in the current scene. From this pool of valid pixels, we randomly select a specified number n_r to serve as the ray samples for the current scene, thereby controlling the overall number of rays

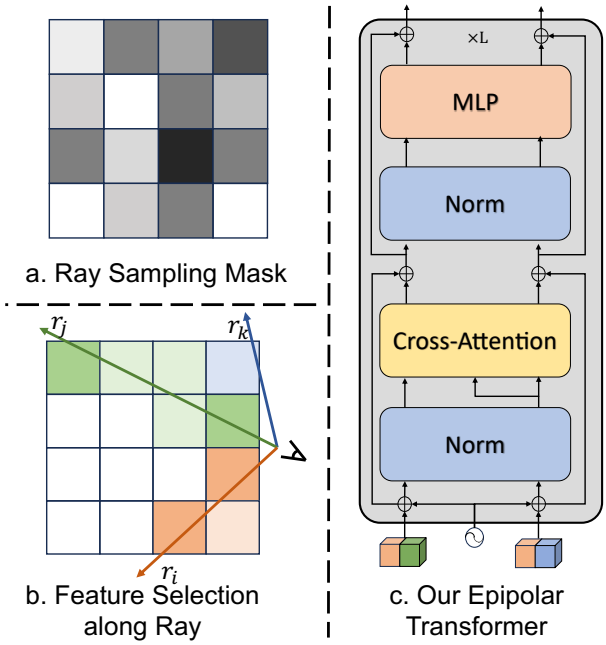

Figure 4. Ray sampling strategy of our proposed ImVoxelENet . In the left figure, we illustrate our ray sampling strategy to select specific number of rays in one image. In the right figure, we demonstrate the strategy to sample feature along a ray.

in the scene.In our approach, we select rays that intersect with more than two points within the 3D voxel space.

In the second module, our objective is to sample voxel features along the already selected rays. After constructing the 3D voxels, we utilize the ray indices obtained from the first step and choose n_r rays that have multiple intersections with the voxels, specifically selecting those rays that have two or more intersection points. For these selected rays, we sample the features along the ray direction by choosing the two outermost voxel features as the sampling result. As illustrated in (b) of Fig. 4, there are typically three scenarios for this process. In the first scenario r_i, where the ray intersects with exactly two voxel points, we directly select these two points. In the second scenario r_k, if the ray intersects with only one voxel, we disregard this ray. In the third scenario r_j, where the ray intersects with multiple voxel points, we select the first and the last voxel points along the ray. It is important to note that the voxel features we select have already undergone multi-view fusion, meaning they incorporate features from all available views. The rationale for this step is that, after fusing multi-view 2D image features, the sampled pixels and rays can effectively capture the corresponding pixel-level feature relationships, thereby establishing accurate epipolar geometric correspondences and achieving implicit 3D representation within the voxel space. The reason for selecting only the first and the last points along the ray is that features that are spatially

close to each other on the same ray might be influenced by noise or random variations during the image capture process, potentially representing similar spatial positions. By focusing on points that are farther apart, we allow the original 3D convolution in the subsequent stages of the network to establish necessary connections. Additionally, by utilizing the transformer structure, we can establish these spatial relationships earlier in the process.

As a summary of this section, our method effectively builds connections between geometric features within the 3D voxel space through ray sampling and along-ray sampling. This approach enhances the network's ability to perceive spatial relationships, ultimately improving its overall performance in 3D object detection tasks.

4.4. Epipolar-Transformer

Next, for the features that have been obtained through sampling and indexing, representing potential geometric relationships, we aim to establish connections between these features to evaluate the correspondence between the two positions along the ray. This involves determining whether these two points share meaningful spatial or geometric relationships that can contribute to a more accurate 3D representation.

To achieve this, we employ an attention-based mechanism, so-called the Epipolar-Transformer structure, which is specifically designed to model these relationships. By inputting the features from the two sampled positions (the frontmost and rearmost voxel points along the ray), the network learns to capture and reinforce the inherent spatial dependencies between them. This process allows the network to identify consistent patterns, structures, and correspondences, thereby facilitating a more accurate representation of the 3D space and enhancing the overall detection performance.

As shown in (c) of Fig. 4, we employ this specific transformer structure, which we refer to as the Epipolar-Transformer. After the sampling process, the input to this module consists of the voxel position's positional embedding combined with the voxel features at the two positions along the ray. These features are obtained by projecting the 2D image features from multiple images onto the voxel space, where overlapping pixels from different images contribute to the same voxel feature. The final feature is represented as the average of all these overlapping 2D image features at each voxel location.

Within the Epipolar-Transformer structure, we design an encoding block centered around a cross-attention mechanism. This encoding block captures the relationships between the sampled features effectively. We stack a total of L such blocks in sequence, allowing the transformer to iteratively refine the representation until the final output is obtained. This design enables the network to comprehensively establish connections between the sampled ray features, enhancing its capability to understand complex spatial and geometric relationships within the 3D space.

5. Experiments

5.1. Experiment Settings

Implement Details Our ImVoxelENet is implemented with PyTorch and MMDetection[4] on Linux workstation armed with Intel E5-2640v4 cpu and 4 Nvidia GTX1080ti graphics cards with 11GB memory. We train our neural network on dataset for 12 epochs using the AdamW optimizer as what have been set in [31]. The initial learning rate is set to 10^{-4}, and it is reduced by a decay factor of 0.1 on the 8th and 11st epoch respectively. We ensure that all other experimental settings closely align with those established for ImVoxelNet baseline. This approach allows for a consistent and fair comparison, minimizing variability and ensuring that any observed differences in performance are attributable to the specific methods being evaluated rather than discrepancies in experimental configurations.

Dataset We evaluate the proposed method on the indoor real scene dataset ScanNet [6]. The dataset is a rich resource for 3D semantic understanding, providing dense 3D reconstructions of indoor scenes along with corresponding RGB-D video sequences. The validation set of ScanNet serves as a robust benchmark for assessing the performance of 3D detection methods, particularly in complex and cluttered environments. This comprehensive dataset comprises 1513 scans that cover over 700 unique indoor scenes. The dataset is divided into a training split, which includes 1201 scans, and a validation split, consisting of 312 scans. Overall, ScanNet contains over 2.5 million images, each accompanied by corresponding depth maps and camera poses. Additionally, the dataset provides reconstructed point clouds with 3D semantic annotations, making it an invaluable resource for 3D object detection and semantic understanding. For the evaluation, we follow the standard protocol established in VoteNet [27]. Specifically, we estimate 3D bounding boxes from the semantic point clouds. The resulting object bounding boxes are axis-aligned, meaning that we do not predict the rotation angle for objects in the ScanNet dataset. This simplification aligns with the common practice in indoor 3D detection, where the primary focus is on the spatial extent and location of objects rather than their orientation.

Metrics We evaluate the comparative methods using the Mean Average Precision (mAP) metric, which serves as a comprehensive measure of detection performance. The mAP metric assesses the average precision across various Intersection over Union (IoU) thresholds, typically ranging

from 0.25 to 0.50 in the context of 3D object detection tasks. This range ensures a balanced evaluation of the model's ability to accurately localize objects at different levels of overlap, providing a more robust and reliable assessment of detection performance across varying degrees of spatial alignment.

5.2. Comparison and Analyse

In Table. 1, we present a comprehensive comparison of our method against other approaches, providing a detailed breakdown of the Average Precision (AP) results across 18 categories within the ScanNetV2 dataset. Additionally, we calculate the mean Average Precision (mAP) at an IoU threshold of 0.25. It is important to note that the methods in the first section of Table 1 utilize 3D point clouds as input to the network, with the final results annotated directly on the 3D point cloud data. In contrast, the second section of the table compares methods that exclusively employ RGB images as input, offering a distinct point of comparison against our approach.

From the results presented in the Table. 1, it can be illustrated that our algorithm outperforms earlier point cloud-based methods by more than 10% in terms of mAP. However, when compared to more advanced point cloud-based algorithms, our approach still exhibits certain limitations. We attribute this to the lack of depth and other three-dimensional geometric information in pure RGB images. As a result, the network is required to implicitly learn the 3D spatial structure and perform epipolar matching to achieve accurate 3D object detection.

Compared to the baseline ImVoxelNet, the performance improvement of our approach stems from the early introduction of epipolar-based foreground-background feature perception modules before the 3D neck CNN module. This strategy enables the model to establish long-range spatial relationships earlier in the process, allowing the network to capture global spatial information more effectively. As a result, this early spatial awareness contributes to the enhanced performance of our network.

We also visualize the comparison in Fig. 5 and Fig. 6. To better illustarte the comparison, we zoom the output of prediction and gt in Fig. 5. In Fig. 6, we randomly select a scene with multiply images as input, and we illustrate the outcome.

5.3. Ablation study

To validate the effectiveness of our ImVoxelENet network structure and the correctness of the integration of its various modules, we conduct comprehensive ablation experiments. All experiments are performed on the ScanNetV2 [6] dataset, using the ImVoxelNet [31] architecture as the baseline network for comparison.

Analysis on number of views. We preprocess the original ScanNet dataset following the default settings of the ImVoxelNet and MMDetection3D frameworks. For each scene, we sample 300 images as multi-view inputs. We train one scene per GPU, and during the training phase, 16 images from each scene are used to update the network parameters. In the testing phase, the default experimental setup involves using 50 images from the validation set as the input for each scene.

To further verify the effectiveness of our proposed framework in establishing robust spatial geometric priors, we conducted additional tests using different numbers of viewpoints and compared the results with our baseline. The specific results are presented in Table. 2. As illustrated, our proposed algorithm consistently achieves superior performance across various numbers of viewpoints, demonstrating its ability to effectively capture spatial relationships. This finding substantiates the effectiveness of our approach in designing and implementing geometric feature connections based on the front-to-back relationships of rays.

Analysis on number of rays. As described earlier, our method samples a varying number of rays from each image and utilizes the features at the intersection points between these rays and the voxels as geometric cues to capture front-to-back relationships. We then design the Ray-Transformer structure to serve as a feature update module, establishing these relationships before the 3D convolution stage. Therefore, it is essential to conduct ablation experiments focused on this module to validate the effectiveness and contribution of our network's design.

Due to computational constraints, we randomly select 1,000 rays from all the available rays across all images as input to the Ray-Transformer in our network design. This specific number is chosen as a balance between computational capabilities and performance. Training with a larger number of rays would require reducing the number of images used per scene to maintain feasibility, which, in turn, would decrease the amount of data the network can utilize. Therefore, in our ablation experiments concerning the number of rays, we fix the number of images per scene and evaluate the impact of varying ray quantities on performance. As previously discussed in earlier sections, we have already analyzed the effect of different numbers of viewpoints on the overall detection performance.

As shown in Table. 3, we present the numerical results of our tests with varying numbers of rays. Intuitively, our network achieves the best performance when using 1,000 rays as the sampling quantity, demonstrating that a greater number of sampled rays leads to improved detection results. However, we acknowledge that this outcome is currently limited by the memory capacity of our computational resources.

Table 1. Quantitative result with multi-view RGB inputs is evaluated on the val-set of ScanNet-V2. The first section of the table presents methods based on point clouds and RGB-D data, while the remaining sections highlight multi-view RGB-only detection approaches.

Methods	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	**mAP@.25**
Seg-Cluster [40]	11.8	13.5	18.9	14.6	13.8	11.1	11.5	11.7	0.0	13.7	12.2	12.4	11.2	18.0	19.5	18.9	16.4	12.2	13.4
Mask R-CNN [9]	15.7	15.4	16.4	16.2	14.9	12.5	11.6	11.8	19.5	13.7	14.4	14.7	21.6	18.5	25.0	24.5	24.5	16.9	17.1
SGPN [40]	20.7	31.5	31.6	40.6	31.9	16.6	15.3	13.6	0.0	17.4	14.1	22.2	0.0	0.0	72.9	52.4	0.0	18.6	22.2
3D-SIS [10]	12.8	63.1	66.0	46.3	26.9	8.0	2.8	2.3	0.0	6.9	33.3	2.5	10.4	12.2	74.5	22.9	58.7	7.1	25.4
3D-SIS (w/ RGB) [10]	19.8	69.7	66.2	71.8	36.1	30.6	10.9	27.3	0.0	10.0	46.9	14.1	53.8	36.0	87.6	43.0	84.3	16.2	40.2
VoteNet [28]	36.3	87.9	88.7	89.6	58.8	47.3	38.1	44.6	7.8	56.1	71.7	47.2	45.4	57.1	94.9	54.7	92.1	37.2	58.7
FCAF3D [30]	57.2	87.0	95.0	92.3	70.3	61.1	60.2	64.5	29.9	64.3	71.5	60.1	52.4	83.9	99.9	84.7	86.6	65.4	71.5
CAGroup3D [36]	60.4	93.0	95.3	92.3	69.9	67.9	63.6	67.3	40.7	77.0	83.9	69.4	65.7	73.0	100.0	79.7	87.0	66.1	75.12
ImVoxelNet	28.5	**84.4**	73.1	70.1	51.9	32.2	**15.0**	34.2	1.6	29.7	66.1	**23.5**	**57.8**	**43.2**	92.4	54.1	**74.0**	**34.9**	48.1
ours	**32.8**	83.9	**73.2**	**72.0**	**52.5**	**33.5**	14.3	**40.5**	**1.9**	**47.8**	**70.4**	20.5	51.8	39.6	**92.6**	**54.2**	72.7	32.5	**49.2**

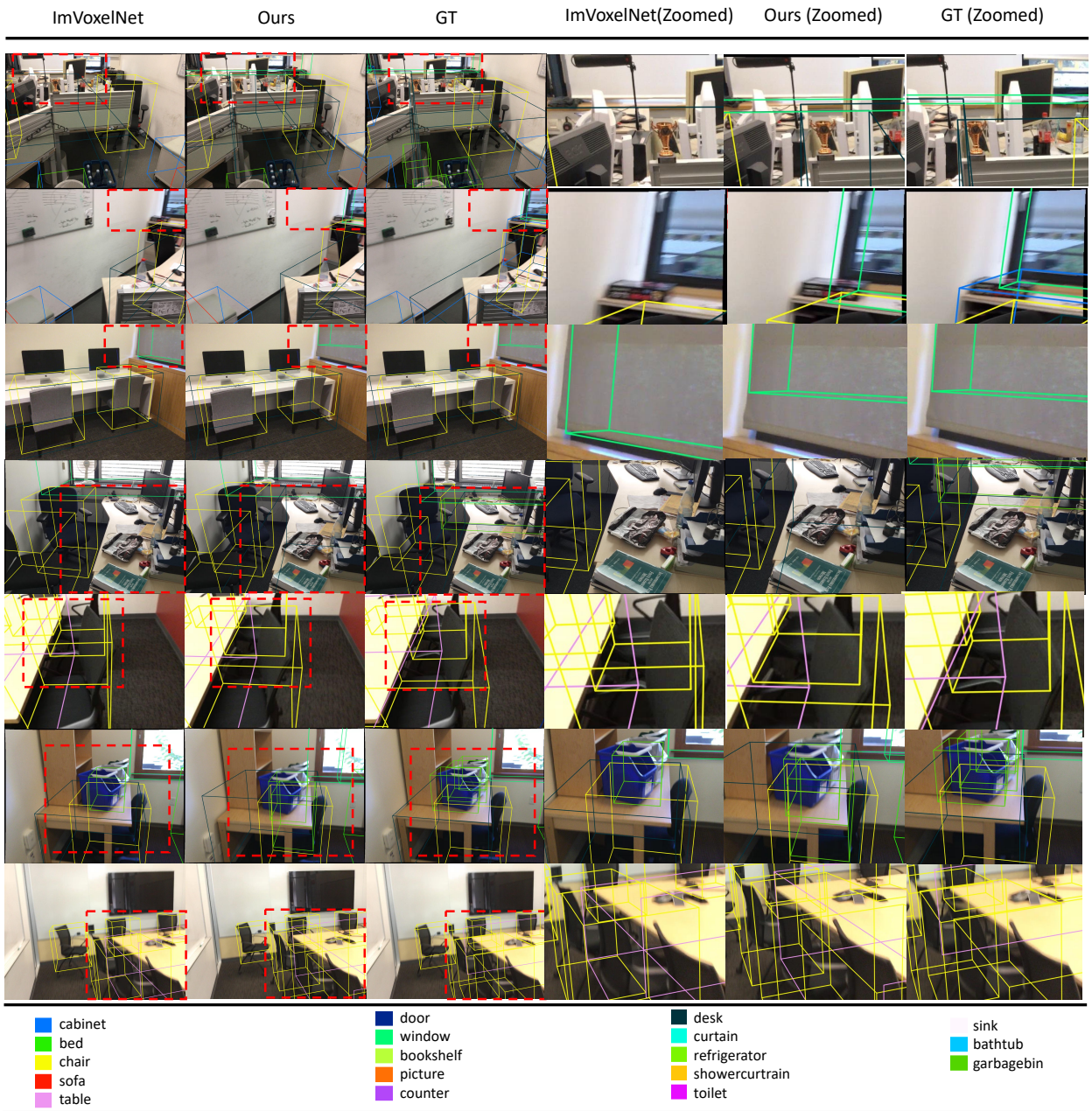

Figure 5. Visual comparison: We compare our ImVoxelENet with the current state-of-the-art method, ImVoxelNet [31], on the ScanNet [6] dataset and visualize a randomly selected image from multiple scenes.

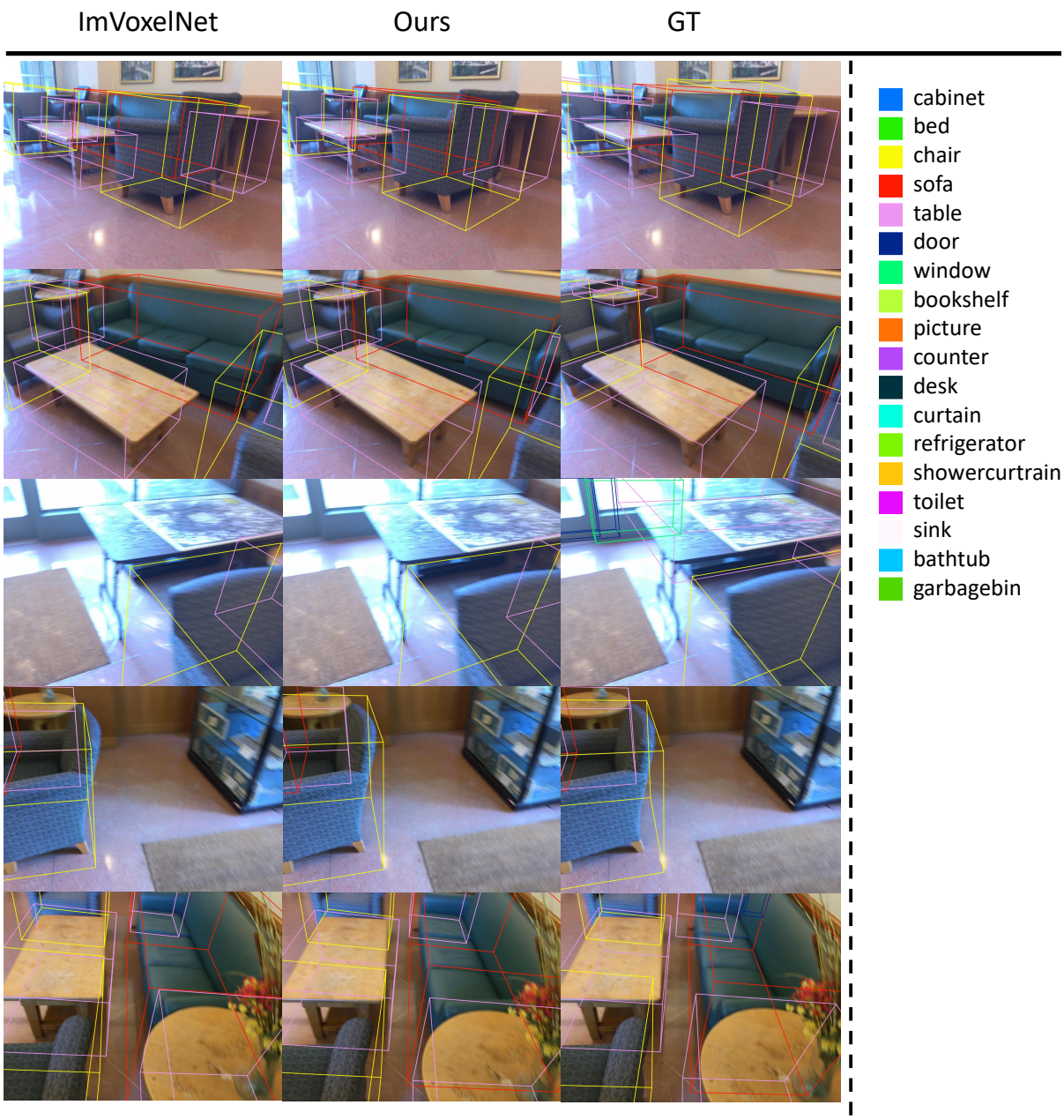

Figure 6. Additional visual comparisons: We also randomly selected several images from a single scene for visual comparisons, providing a comprehensive view of the scene.

Analysis on modules. Next, we investigate the impact of different module parameters and structures on the network's performance. We start by examining how varying the number of attention layers influences the final results. The number of blocks directly affects the total parameters in our ImVoxelENet network, which plays a critical role in updating the voxel grid features. To evaluate this, we adjusted this hyperparameter, and the results indicate that a 6-layer attention structure provides optimal performance. As illustrated in Table. 4, this finding demonstrates the necessity of adequately capturing spatial attention across the front-to-back geometric space, as it significantly enhances the network's ability to reflect the actual 3D detection capabilities.

In the final set of ablation experiments, we designed a module replacement test. In this experiment, we replaced the cross-attention module within the Ray-Transformer structure with a Multi-Layer Perceptron (MLP) module. During this replacement process, we also performed a re-

Table 2. Ablation results: We compared the detection results using different numbers of input images. The upper section presents the results of the baseline method, while the lower section shows the results of our proposed method.

id	nums of views	mAP@0.25↑	mAP@0.5 ↑
1	10	38.9	16.4
2	20	41.4	19.2
3	30	44.8	20.9
4	40	47.3	22.3
5	50	48.1	23.1
1	10	40.2	17.2
2	20	43.0	20.5
3	30	46.6	22.9
4	40	48.9	23.9
5	50	49.2	24.2

Table 3. Ablation results: We compared the results of our proposed method under different numbers of rays as input.

id	nums of rays	mAP@0.25↑	mAP@0.5↑
1	0	48.1	23.1
2	500	48.3	23.7
3	1000	49.2	24.2

Table 4. Ablation results: We compared the results of our proposed method using different numbers of layers.

id	num of layers	mAP@0.25↑	mAP@0.5↑
1	2	48.3	23.6
2	4	48.9	24.1
3	6	49.2	24.2

Table 5. Ablation on different blocks in ray-transformer. We test the different block by replacing the cross attention(CA) into other architecture. MLP represent that we replace the CA block with mlp layers.

id	block type	mAP@0.25↑	mAP@0.5↑
1	CA	49.2	24.2
2	MLP	48.2	22.7

shape operation on the corresponding features to ensure compatibility with the modified network architecture. The experimental results, as shown in Table. 5, indicate that while the MLP module does provide some performance improvement, the degree of enhancement is significantly less than that achieved by the original cross-attention mechanism. This outcome further validates the effectiveness of our proposed algorithm and network structure, demonstrating that the cross-attention approach is more adept at capturing the necessary geometric relationships for optimal 3D object detection.

6. Conclusion

In this paper, we introduce a novel deep learning-based network framework aimed at enhancing 3D object detection performance in indoor scenes. We conduct a thorough analysis of the limitations and challenges of previous methods, identifying that a key issue in current multi-view object detection algorithms lies in the inaccurate epipolar matching across different views, which hinders both performance and efficiency.

To address this problem, we start from the back-projection stage and redesign the process of establishing connections between voxel and pixel features. Specifically, we develop a ray-based sampling method that captures long-range geometric relationships beyond the original voxels by considering the foreground-background relationships along each ray. Building on this foundation, we implement an attention mechanism that leverages these relationships, enabling the network to effectively link features along the ray and thereby enhance its implicit understanding of 3D spatial geometry.

As a result, the proposed network achieves significant improvements in detection performance. We further validate the effectiveness and efficiency of our approach through comprehensive comparative and ablation studies, demonstrating its superiority over existing methods.

Limitation However, we also clearly acknowledge that our proposed network still has certain limitations. While our approach demonstrates performance improvements in multi-view detection tasks within indoor scenes, these environments typically benefit from stronger structural and categorical priors. In contrast, outdoor scenes present greater challenges due to their increased variability and diversity of object types, which our current network is less equipped to handle effectively. Addressing this gap will require more targeted designs and implementations. In future work, we aim to develop strategies that better adapt to the complexities of outdoor scenarios, enhancing the robustness and generalizability of our method.

Acknowledgement

This work is funded by the National Natural Science Foundation of China under Grant 61972014.

References

[1] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[2] S. Bi, Z. Xu, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Deep reflectance vol-

umes: Relightable reconstructions from multi-view photometric images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 294–311. Springer, 2020. 2

[3] G. Brazil, A. Kumar, J. Straub, N. Ravi, J. Johnson, and G. Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 2

[4] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 8

[5] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez. Focalformer3d: Focusing on hard instance for 3d object detection. 2023. 3

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 8, 9, 10

[7] J. M. Fácil, A. Concha, L. Montesano, and J. Civera. Single-view and multi-view depth fusion. *IEEE Robotics and Automation Letters*, 2(4):1994–2001, 2017. 2

[8] C. Han, J. Yang, J. Sun, Z. Ge, R. Dong, H. Zhou, W. Mao, Y. Peng, and X. Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics and Automation Letters*, 2024. 3

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 10

[10] J. Hou, A. Dai, and M. Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *CVPR*, 2019. 10

[11] S. Huang, S. Qi, Y. Xiao, Y. Zhu, Y. N. Wu, and S.-C. Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *Advances in Neural Information Processing Systems*, 31, 2018. 1, 4

[12] X. Jiang, S. Li, Y. Liu, S. Wang, F. Jia, T. Wang, L. Han, and X. Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. *arXiv preprint arXiv:2308.09616*, 2023. 3, 4

[13] R. L. J. Z. Z. Z. Y. C. Jinglin Zhan, Tiejun Liu. Real-aug: Realistic scene synthesis for lidar augmentation in 3d object detection. 2023. 3

[14] M. Kölle, D. Laupheimer, S. Schmohl, N. Haala, F. Rottensteiner, J. D. Wegner, and H. Ledoux. The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1:100001, 2021. 2

[15] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 3, 4, 6

[16] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4865–4874, 2017. 1

[17] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, H. Tian, E. Xie, J. Xie, L. Chen, T. Li, Y. Li, Y. Gao, X. Jia, S. Liu, J. Shi, D. Lin, and Y. Qiao. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2023. 3

[18] F. Liu, T. Huang, Q. Zhang, H. Yao, C. Zhang, F. Wan, Q. Ye, and Y. Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. *arXiv preprint arXiv:2402.03634*, 2024. 3

[19] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 3

[20] J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2723–2732, 2021. 1

[21] J. Mao, S. Shi, X. Wang, and H. Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023. 4

[22] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 2, 4, 5, 6, 7

[23] M. Naseer, S. Khan, and F. Porikli. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access*, 7:1859–1887, 2018. 1

[24] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, and J. J. Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 4

[25] J. Ong, B.-T. Vo, B.-N. Vo, D. Y. Kim, and S. Nordholm. A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2246–2263, 2020. 2

[26] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 3

[27] C. R. Qi, O. Litany, K. He, and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 3, 8

[28] C. R. Qi, O. Litany, K. He, and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 10

[29] S. Qi, X. Ning, G. Yang, L. Zhang, P. Long, W. Cai, and W. Li. Review of multi-view 3d object recognition methods based on deep learning. *Displays*, 69:102053, 2021. 2

[30] D. Rukhovich, A. Vorontsova, and A. Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *Eu-*

ropean Conference on Computer Vision, pages 477–493. Springer, 2022. 10

[31] D. Rukhovich, A. Vorontsova, and A. Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 2, 4, 5, 6, 7, 8, 9, 10

[32] S. Stekovic, M. Rad, A. Moradi, F. Fraundorfer, and V. Lepetit. Mcts with refinement for proposals selection games in scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[33] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1

[34] R. Tucker and N. Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 4

[35] B.-H. Wang, F. Wijaya, R. Fischer, Y.-H. Tang, S.-J. Wang, W.-E. Hsu, and L.-C. Fu. A scene understanding and positioning system from rgb images for tele-meeting application in augmented reality. In *2023 9th International Conference on Virtual Reality (ICVR)*, pages 106–114. IEEE, 2023. 4

[36] H. Wang, L. Ding, S. Dong, S. Shi, A. Li, J. Li, Z. Li, and L. Wang. CAGroup3d: Class-aware grouping for 3d object detection on point clouds. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 10

[37] Q. Wang, Y.-Y. Chang, R. Cai, Z. Li, B. Hariharan, A. Holynski, and N. Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 5

[38] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 3

[39] T. Wang, H. Sheng, R. Chen, R. Cong, M. Zhao, and Z. Cui. Adaptive epi-matching cost for light field disparity estimation. *IEEE Transactions on Instrumentation and Measurement*, 2024. 4

[40] W. Wang, R. Yu, Q. Huang, and U. Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *CVPR*, 2018. 10

[41] F. Williams, T. Schneider, C. Silva, D. Zorin, J. Bruna, and D. Panozzo. Deep geometric prior for surface reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10130–10139, 2019. 2

[42] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020. 4

[43] K. Xiong, S. Gong, X. Ye, X. Tan, J. Wan, E. Ding, J. Wang, and X. Bai. Cape: Camera view position embedding for multi-view 3d object detection. 2023. 3

[44] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3, 4

[45] C. Yang, J. Zheng, X. Dai, R. Tang, Y. Ma, and X. Yuan. Learning to reconstruct 3d non-cuboid room layout from a single rgb image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2534–2543, 2022. 1

[46] T. Yin, X. Zhou, and P. Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1

[47] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2

[48] Z. Yu, Q. Liu, W. Wang, L. Zhang, and X. Zhao. Polarbevdet: Exploring polar representation for multi-view 3d object detection in bird's-eye-view. *arXiv preprint arXiv:2408.16200*, 2024. 3, 4

[49] C. Zhang, Z. Cui, C. Chen, S. Liu, B. Zeng, H. Bao, and Y. Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021. 2

[50] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, and S. Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 1, 4

[51] D. Zhang, G. Wang, R. Zhu, J. Zhao, X. Chen, S. Zhang, J. Gong, Q. Zhou, W. Zhang, N. Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 3

[52] D. Zhang, Z. Zheng, H. Niu, X. Wang, and X. Liu. Fully sparse transformer 3d detector for lidar point cloud. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[53] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5287–5295, 2017. 1

[54] Z. Zhang, F. Cole, R. Tucker, W. T. Freeman, and T. Dekel. Consistent depth of moving objects in video. *ACM Transactions on Graphics (ToG)*, 40(4):1–12, 2021. 5

[55] Z. Zhang, B. Sun, H. Yang, and Q. Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. 3, 6

[56] Q. Zhou, J. Cao, H. Leng, Y. Yin, Y. Kun, and R. Zimmermann. Sogdet: Semantic-occupancy guided multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7668–7676, 2024. 1

[57] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022. 2

[58] Z. Zong, D. Jiang, G. Song, Z. Xue, J. Su, H. Li, and Y. Liu. Temporal enhanced training of multi-view 3d object detector via historical object prediction, 2023. 3