SFMNet: Sparse Focal Modulation for 3D Object Detection

Oren Shrout
Technion, Israel
shrout.oren@campus.technion.ac.il

Ayellet Tal
Technion, Israel
ayellet@ee.technion.ac.il

https://github.com/oshrout/SFMNet

Abstract

We propose SFMNet, a novel 3D sparse detector that combines the efficiency of sparse convolutions with the ability to model long-range dependencies. While traditional sparse convolution techniques efficiently capture local structures, they struggle with modeling long-range relationships. However, these relationships are essential for 3D object detection. In contrast, transformers are designed to capture these long-range dependencies through attention mechanisms. But, they come with high computational costs, due to their quadratic query-key-value interactions. Furthermore, directly applying attention to non-empty voxels is inefficient due to the sparse nature of 3D scenes. Our SFMNet is built on a novel Sparse Focal Modulation (SFM) module, which integrates short- and long-range contexts with linear complexity by leveraging a new hierarchical sparse convolution design. This approach enables SFMNet to achieve high detection performance with improved efficiency, making it well-suited for large-scale LiDAR scenes. We show that our detector achieves state-of-the-art performance on autonomous driving datasets.

1. Introduction

3D object detection is fundamental for applications in autonomous driving [2, 3, 17, 30, 43, 53] and robotics [46, 69]. Extensive research has explored various methods, including point-based [33–37, 39, 40, 57], range-based [13, 44, 47], and voxel-based approaches [5, 7, 10, 23, 38, 41, 42, 49, 55, 58, 59, 61, 62, 64–66]. Point-based methods process raw point clouds but often suffer from high computational costs or reduced accuracy in large-scale scenes due to point sampling. Range-based methods project 3D points to apply 2D convolutions, making them computationally efficient. However, they often struggle with occlusions, object localization, and size regression errors, as they lack the expressive 3D features necessary for accurate detection.

Voxel-based methods, which are the focus of this paper, transform irregular point clouds into regular voxel grids. This enables the use of convolutions, specifically sparse convolutions, that operate efficiently despite the presence of many empty voxels [9, 20, 55]. While effective at capturing local structures, methods based on submanifold sparse convolution [19] substantially limit the receptive field, thereby

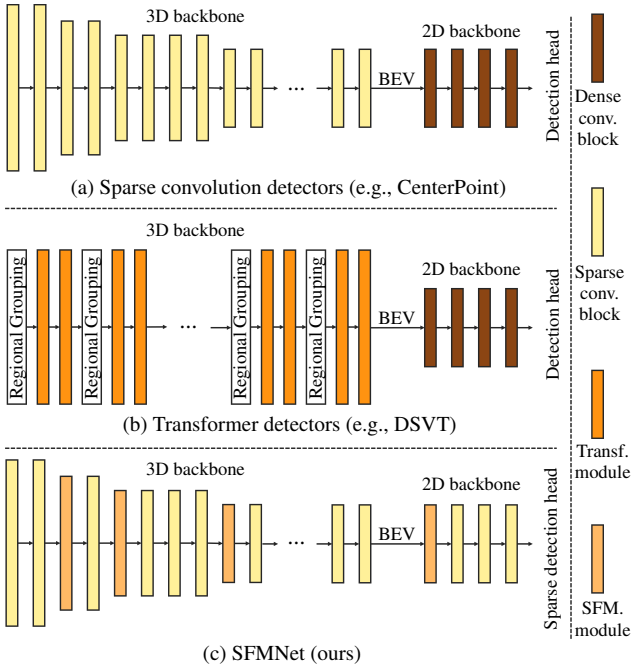


Figure 1. **SFMNet in comparison with other architectures.** Similar to sparse convolution-based detectors, e.g.,[59] (a), our SFMNet detector (c) comprises a multi-stage 3D backbone with sparse convolution blocks. In contrast, SFMNet is fully sparse, including the 2D backbone, and integrates a novel SFM module (light orange) into both the 3D and 2D backbones. Furthermore, unlike transformers, SFMNet uses sparse convolutions as the token mixer, which makes query modulation efficient. Moreover, unlike the single-stride network design used in transformer-based detectors (e.g., [49]), our SFMNet detector employs an encoder-based design.

reducing the model's ability to capture long-range dependencies, which are critical for detection [52]. In 3D detection, long-range dependencies encompass relationships among voxels within an object as well as between the object and its immediate surroundings. Given the inherent sparsity and frequent occlusions in LiDAR data, relying only on local features makes detecting small objects, such as pedestrians, extremely difficult when only a few points are available. Contextual information, such as the presence of a crosswalk, can greatly aid detection, as even a few points above it may indicate a pedestrian's presence, highlighting the importance of long-range dependencies.

To address these limitations, the use of transformers has been proposed [8, 14, 45, 48–50, 68]. With self-attention (SA) as their core mechanism, transformers can capture global relationships, increasing the detector representation capacity. Despite its benefits, the quadratic complexity of SA is a significant bottleneck, particularly in autonomous vehicle scenarios where LiDAR scenes span hundreds of meters and contain hundreds of thousands of points.

We propose a 'compromise' between the two approaches by using large kernels in convolutional neural networks. This strategy has demonstrated efficiency, primarily in 2D [11, 12, 26]. Large-kernel methods improve representation capacity by increasing the receptive field and offering greater robustness compared to transformers [4]. However, this approach is often impractical for 3D scenes, as the number of parameters grows cubically, necessitating techniques like sophisticated kernel weight partitioning with shared weights [6], data-driven kernel generation [28], or weight pruning [16]. Although these methods reduce parameter bottlenecks and enlarge the effective receptive field, they offer minimal accuracy improvements and are less efficient than traditional sparse convolutions.

Inspired by these solutions, this paper aims to combine the best of both worlds—enabling transformer-like long-range dependencies while maintaining the efficiency of sparse convolutions. We introduce a new 3D sparse detector, termed *SFMNet*, that relies entirely on (sparse) convolutions; however, between these convolutions, we introduce a novel 'transformer-like' module where the attention mechanism is also replaced with convolutions. Our *SFM* module captures both short- and long-range relationships between voxels, bridging gaps left by traditional sparse convolution detectors while providing greater efficiency than transformer-based detectors. As illustrated in Fig. 1, our model differs from traditional sparse convolution detectors by being fully sparse, including its 2D backbone, and by incorporating our SFM module within the backbones.

The *Sparse Focal Modulation (SFM)* module modulates short- and long-range contexts and integrates them with query tokens, similar to transformers. However, instead of the quadratic complexity of self-attention, SFM operates with linear complexity relative to the number of tokens (voxels). The core idea is to efficiently extract local contexts using sparse convolutions in a hierarchical structure, enabling the design to capture global relationships. Limiting attention to local context windows has been shown to reduce the complexity bottleneck [27]. However, due to the sparse nature of LiDAR points, the number of occupied voxels in each local window can vary significantly, making efficient local-based attention non-trivial [49]. To efficiently extract context windows, we design our SFM module with stacked submanifold sparse convolutions, using small kernels and varying dilation rates. This design enables hierarchical context aggregation, progressively expanding the receptive field at each level. Sparse convolutions capture local structures efficiently, while the hierarchical design and increased dilations allow for learning sparse long-range dependencies.

We validate the effectiveness of our detector on three well-known autonomous driving datasets. Our results demonstrate that the network advances the state-of-the-art in long-range detection (Argoverse2 [53]), for which it is designed. On mid- to short-range datasets, Waymo Open [43] and nuScenes [2], it performs comparably or slightly better than previous state-of-the-art detectors.

Thus, the main contributions are twofold:

- We propose a novel sparse focal modulation (SFM) module that enables long-range dependencies like transformers while maintaining the efficiency of sparse convolutions. It hierarchically aggregates contextual information from sparse inputs, outperforming traditional attention methods by reducing computational overhead and efficiently handling sparse voxels of varying densities.
- We introduce SFMNet, a 3D object detector built on top of the SFM module, which effectively handles large point cloud scenes. We validate the detector's usefulness on well-known large-scale LiDAR datasets, achieving state-of-the-art results.

2. Related work

3D detection on point clouds. 3D detection methods are broadly categorized into point-based methods and voxel-based methods. Point-based methods use PointNet [33, 34]-like architectures to extract geometric features directly from raw point clouds [33–37, 39, 40, 51, 57]. However, in large-scale point cloud scenes typical of outdoor LiDAR datasets in autonomous driving, these methods suffer from lower accuracy and high processing times due to point sampling and neighbor searching. In contrast, voxel-based methods convert irregular point clouds into regular voxel grids and apply sparse convolutions on non-empty voxels [5, 7, 10, 23, 25, 38, 41, 55, 58, 59, 61, 62, 64–66]. While efficient for processing large-scale data, these methods rely on submanifold sparse convolution, which might limit network representation capacity by prioritizing effi-

ciency over capturing long-range dependencies.

Long-range dependencies for 3D detection. Recent 3D detection methods capture long-range dependencies using large-kernel sparse convolutions or transformers. Large-kernel convolution approaches expand the receptive field by increasing kernel sizes, a technique that has shown accuracy gains in 2D detection [11, 12, 26]. However, directly applying large kernels to 3D scenes incurs a cubically increasing computational overhead. To address this, LargeKernel3D [6] introduces a spatial partitioning convolution with shared weights among neighboring voxels. LinK [28] further optimizes efficiency by generating sparse kernels dynamically, assigning weights only to non-empty voxels rather than maintaining a dense kernel matrix. Similarly, LSK3DNet [16] reduces computation by pruning redundant weights across spatial and channel dimensions. While these methods alleviate some computational demands, they still face performance challenges in large-scale 3D data.

Transformer-based detectors, by contrast, use self-attention to capture long-range dependencies, enabling global token interactions. However, applying attention in high-resolution 3D scenes with many tokens is computationally prohibitive [8, 50, 68]. To mitigate this, VoTr [31] implements local and dilated attention, focusing on sparse voxel interactions. SWFormer [45] utilizes a window-based transformer with a bucketing scheme to manage windows of varying non-empty voxels. SST [14] refines token processing by grouping tokens regionally, applying attention within each group, and shifting regions to capture structured interactions. DSVT [49] sequentially applies windowed self-attention along the X and Y axes to capture dependencies with reduced complexity. Although they improve long-range relationships, they still face quadratic scaling issues with context-window tokens due to self-attention.

We propose a network that combines the strengths of sparse convolutions and transformers. Our design maintains the efficiency of sparse convolutions while enabling long-range dependencies similar to transformers. However, unlike traditional transformers, our module is built from sparse convolutions and operates with linear complexity relative to the number of tokens (voxels). Furthermore, eliminating the need for sampling helps us achieve better accuracy.

3. Method

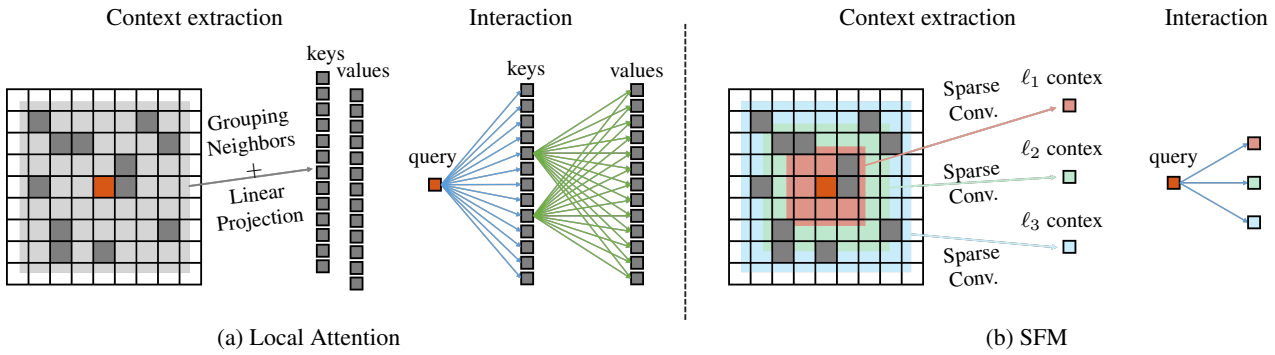
Voxel-based detectors are widely used for 3D detection in outdoor scenes, such as those in autonomous driving scenarios. They commonly leverage sparse convolutions for efficient learning in sparse environments. There are two main types of sparse convolutions: regular [18] and submanifold [19]. Regular sparse convolutions expand features into (not necessarily occupied) neighboring voxels, enabling information exchange between spatially disconnected regions, but increasing feature map density. Their expansive nature makes them suitable for downsampling and upsampling. Submanifold sparse convolutions operate only on occupied voxels, ensuring that the output retains the same sparsity as the input, thereby facilitating efficient learning of local structures. However, due to their limited kernel size, they lack long-range dependencies, which are useful for object detection [52].

We aim to design an efficient detector built from sparse convolutions capable of capturing both short- and long-range dependencies. While transformers are effective for modeling global relationships in 2D, applying them to sparse point clouds is challenging: (1) With numerous occupied voxels in each scene, standard self-attention suffers from massive query-key and query-value interactions, leading to a high computational overhead or suboptimal accuracy due to subsampling [32, 63]. (2) The varying voxel density across regions hinders efficient batching [14, 49].

To address both the computational overhead and the varying number of occupied voxels within a context window, we propose replacing self-attention with *Sparse Focal Modulation (SFM)* as the token mixer in the transformer. The SFM module reduces query interactions to only a few focal contexts around each query by hierarchically extracting representations of the surrounding connections. Additionally, sparse convolutions were designed to efficiently handle varying numbers of occupied voxels within each local region and can facilitate effective context extraction.

Our SFM module is described in Section 3.1. Fig. 2 illustrates its context extraction and interaction phases compared to naive local attention. Local attention involves extracting a context for each query token, then performing query-key and query-value interactions. This requires identifying all non-empty voxels in the context window, sampling or padding them to a fixed maximum, and concatenating features for batch processing. In high-resolution scenes with hundreds of thousands of voxels, this quickly becomes a bottleneck, and query-key-value interactions incur quadratic complexity with the number of tokens per window. In contrast, SFM modulates query interactions by extracting multiple context levels via sparse convolutions, capturing short-range structures at lower levels and long-range dependencies at higher levels. These sparse convolutions efficiently handle varying numbers of non-empty voxels within context windows, enabling effective batch processing. By using convolutions with different receptive fields for context extraction, the interaction phase is reduced to only the number of context levels.

The effectiveness of our proposed SFM is demonstrated through a novel sparse detector, called *SFMNet,* introduced in Section 3.2. Our detector effectively captures long-range dependencies with a large receptive field, improving performance in object detection tasks.



(a) Local Attention (b) SFM

Figure 2. **SFM module vs. Local attention.** Given a query token (burnt orange), local-based attention (a) identifies and groups non-empty tokens (voxels) within a context window (light gray) to form key and value tokens. The query-key-value interactions have quadratic complexity. In contrast, SFM (b) leverages sparse convolutions to extract multiple levels of context, reducing query interactions to only a few focal contexts and achieving linear complexity with respect to the number of non-empty voxels in the context window. Furthermore, while local-based attention suffers from inefficiencies in batch processing due to the varying numbers of non-empty voxels within each context window, our SFM approach is robust to this variability, extracting a constant number of contexts for each window.

3.1. Sparse focal modulation

Focal Modulation (FM) is an alternative token mixer to self-attention, designed to limit token interactions. It has been applied in 2D tasks like classification, segmentation, and detection [29, 56]. Extending FM to 3D faces two primary challenges. First, 3D representations are inherently sparse, with voxel count scaling cubically. For instance, a typical Argoverse2 [53] scene spans $400 \times 400 \times 8$ meters, yielding approximately 640 million voxels at $0.1 \times 0.1 \times 0.2$ m resolution. With only about 100,000 sampled points, this results in a high percentage of empty voxels. Second, 2D FM typically captures global context using stacked depth-wise convolutions with progressively larger kernels to model multi-scale visual features. However, in 3D detection, sparse depth-wise convolutions offer no benefit over spatial ones and are memory-intensive [6]. Furthermore, applying large kernels in 3D scenes leads to a cubic increase in parameter count, substantially raising computational costs.

We propose a new *Sparse Focal Modulation (SFM)* module that addresses these challenges and captures both short- and long-range contexts while adaptively exchanging information with each token query. Unlike the conventional approach of using large kernels to expand the context window densely, we expand the context window sparsely with small kernels and increased dilation rates. We reason that queries related to small objects often rely on local structures and benefit from compact, focused context windows. In contrast, queries for large objects (e.g., partially occluded vehicles or roads) require broader but not necessarily dense context windows, given the sparsity of LiDAR point clouds. Hence, our context aggregation is sparse.

Fig. 3 illustrates our SFM module, which modulates each non-empty voxel feature (x) with surrounding information

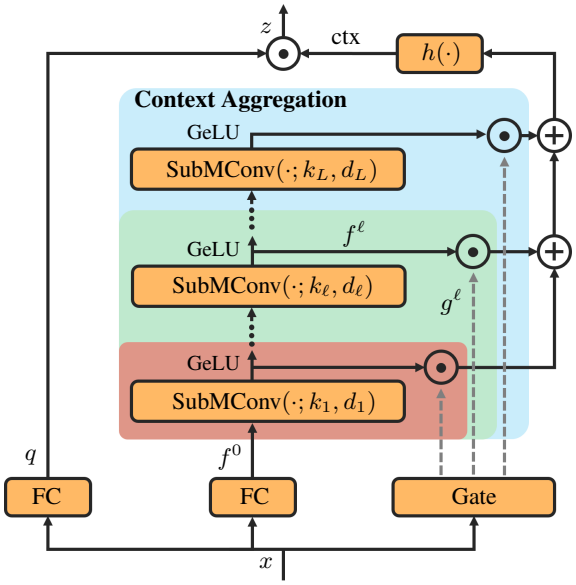


Figure 3. **Sparse Focal Modulation (SFM).** SFM leverages sparse convolutions to aggregate multi-level contexts and modulate them with query features, efficiently capturing both local and global dependencies.

to produce a spatially aware feature (z). Its core component, the *Context Aggregation (CA)* block, captures context for each query token by: 1) efficiently extracting contexts hierarchically, from local to global regions (e.g., red, green, and blue regions in Fig. 2); and 2) adaptively aggregating context features at each level using a gating mechanism to provide meaningful context for each query. The aggregated context of the CA block, ctx, is modulated with the corresponding query q to generate the feature z.

Submanifold sparse convolutions efficiently handle voxel sparsity, as they operate strictly on non-empty voxels. By stacking submanifold sparse convolutions with small kernels, we can effectively capture local contexts at each level. Specifically, given a feature vector $x \in \mathbb{R}^{N\times C}$, where N represents the number of voxels and C denotes the number of features, we project this vector into new feature spaces using linear layers to generate both per-voxel queries $q \in \mathbb{R}^{N\times C}$ and initial focal features $f^0 \in \mathbb{R}^{N\times C}$. The focal features are then fed into the Context Aggregation (CA) module. The CA produces L levels of focal feature maps, f^ℓ, with each level fed to a submanifold sparse filter that extracts the next focal context. Each level has a progressively larger effective receptive field than the previous level. At level ℓ, the focal feature $f^\ell \in \mathbb{R}^{N\times C}$ is computed as

$$f^\ell = \text{GeLU}(\text{SubMConv}(f^{\ell-1}; k^\ell, d^\ell)), \tag{1}$$

where $\text{SubMConv}(\cdot; k^\ell, d^\ell)$ denotes a submanifold sparse convolution with kernel size k^ℓ and dilation rate d^ℓ, and GeLU is an activation function [21]. The hierarchical convolutions effectively capture multiple levels of focal abstraction, with the effective receptive field at level ℓ given by:

$$r^\ell = 1 + \sum_{i=1}^{\ell} (k^i - 1) \cdot d^i. \tag{2}$$

When aggregating context features at each level, we note that not all queries benefit equally from short-range or long-range contexts. Specifically, queries for small objects are more likely to gain from the short-range context available at lower focal levels, while larger objects or background regions tend to benefit more from the long-range context captured at higher focal levels. To address this, a gating mechanism is employed to weigh the focal features from the different levels. This gate adaptively regulates the amount of context extracted from each level, enabling the CA module to learn suitable contexts for various queries. By combining the focal features with the gate weights, the CA produces a context feature vector for each query, defined as:

$$\text{ctx} = \text{h}(\sum_{\ell=1}^{L} f^\ell \cdot g^\ell) \in \mathbb{R}^{N\times C}, \tag{3}$$

where $g^\ell \in \mathbb{R}^{N\times 1}$ represents the gate weight for focal level ℓ, and $h(\cdot)$ is a linear layer that projects the output context into the query space.

Finally, we modulate the focal contexts (ctx) to the queries (q) through element-wise multiplication, which enables each query to learn its relevant context:

$$z = q \cdot \text{ctx} \in \mathbb{R}^{N\times C}. \tag{4}$$

In summary, to address voxel sparsity in 3D space, we propose Sparse Focal Modulation (SFM), which expands the context window efficiently using submanifold sparse convolutions with small kernels and increasing dilation. To handle varying contextual needs, the Context Aggregation (CA) block employs a gating mechanism to adaptively aggregate multi-scale features, allowing each query to focus on the most relevant context for accurate 3D detection.

To demonstrate its utility, we incorporated a single SFM module into each backbone (3D and 2D) within the CenterPoint [59] architecture, without otherwise altering the structure. This has already yielded a 2.6% improvement in Level-2 mAPH on a 20% subset of the Waymo Open Dataset [43].

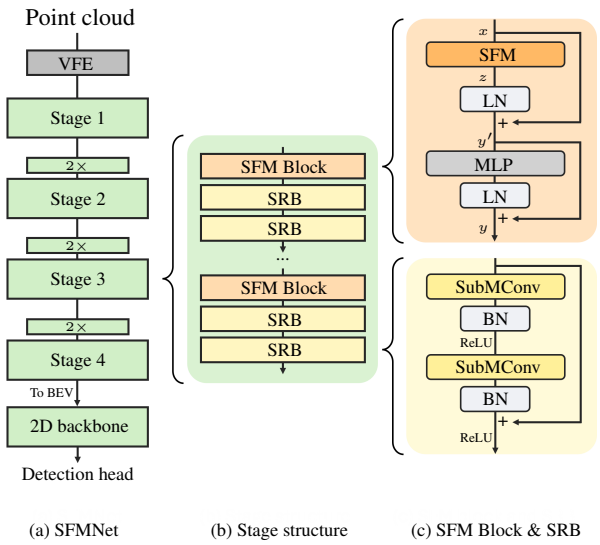
3.2. SFMNet

Building on our sparse focal modulation (SFM) module, we present SFMNet, a simple and efficient sparse 3D object detector. It enables learning long-range dependencies that convolution-based backbones lack, while maintaining sparse convolution efficiency.

Our architecture is built on two core principles. First, increasing the effective receptive field (ERF) in mid- and high-level layers enhances the network's ability to capture broader spatial context—this is primarily achieved using a single SFM block. Second, once a sufficient receptive field is established, adding depth with SRB blocks becomes more efficient, as they act like small kernels and offer better parameter utilization. In practice, we find that one SFM block combined with one SRB block is sufficient to meet our design goals. When further deepening the network, we prioritize adding more SRB blocks rather than additional SFM blocks (see Section 4.2). This design choice aligns with the observation from [12] that, after the receptive field is adequately covered, small kernels become more efficient than large ones—an analogy that maps well to our use of SFM (large kernel equivalent) and SRB (small kernel equivalent) blocks.

Fig. 4(a) illustrates the SFMNet architecture. The input point cloud is voxelized, and features are extracted using a dynamic voxel feature extraction (VFE) method [67]. These features are then processed by our 3D and 2D sparse backbones. Our 3D backbone includes four stages and between each pair of consecutive stages, a regular sparse convolution with stride 2 is used for downsampling. Each stage includes an SFM block followed by multiple submanifold residual blocks (SRBs) for high-level feature extraction (Fig. 4(b)). We use a variable number of SFM blocks and SRBs to enhance representational capacity and increase depth depending on the dataset. The higher the stage, the more SRB blocks per SFM block should be added to scale the model.

The SFM block employs our SFM module as the token mixer within a standard MetaFormer-like design [60], as illustrated in Fig. 4(c), where non-empty voxels serve as the tokens. Formally, given an input feature map $x \in \mathbb{R}^{N\times C}$



(a) SFMNet (b) Stage structure (c) SFM Block & SRB

Figure 4. **SFMNet architecture.** (a) The SFMNet architecture extracts voxel features from a point cloud input (VFE), processes them through multiple stages, and converts them to a bird's-eye view (BEV) representation before passing the output to a 2D backbone and detection head. (b) Each stage consists of multiple Sparse Focal Modulation (SFM) blocks followed by Sparse Residual Blocks (SRBs). (c) The SFM block aggregates contextual information and processes it, while the SRB employs SubMConv layers with batch norm (BN) and ReLU activations to capture spatial features.

and its corresponding focal-modulated feature $z \in \mathbb{R}^{N \times C}$ generated by the SFM module, as described in Section 3.1, the SFM block is formulated as:

$$y' = \text{LN}(z) + x, \quad y = \text{LN}(\text{MLP}(y')) + y', \tag{5}$$

where LN denotes LayerNorm, the multilayer perceptron (MLP) has a single hidden layer, and y is the refined feature from focal sparse modulation. The 3D sparse features are then compressed into 2D bird's-eye view (BEV) features, as in [61], and passed to a 2D backbone. Like the 3D backbone, the 2D backbone includes an SFM block followed by SRBs, but uses 2D sparse convolutions.

4. Experiments

Datasets and metrics. To validate our approach's effectiveness, we conducted experiments on three popular datasets: Argoverse2 [53], Waymo Open [43], and nuScenes [2]. The detection range of these datasets varies, with long-range detection in Argoverse2 (up to ± 200 meters), mid-range in Waymo Open (up to ± 75 meters), and short-range in nuScenes (up to ± 54 meters). We use the standard evaluation metrics for each dataset: mAP for Argoverse2, mAP and mAP weighted by heading accuracy (mAPH) for Waymo Open, and mAP and nuScenes detection score (NDS) for nuScenes. On Waymo, the metrics are applied to two difficulty levels: L1 for objects with more than five LiDAR points and L2 for those with at least one sampled point.

Implementations details. To compare with prior state-of-the-art methods, we trained SFMNet for 24 epochs on Argoverse2 and Waymo Open, and 20 epochs on nuScenes. Ablation studies used 6 epochs on Argoverse2. Additional implementation details are provided in the supplemental.

4.1. Results

Quantitative results. We expect our model to be particularly beneficial for long-range scenes. Indeed, as shown in Table 1, which compares our method to state-of-the-art sparse convolution-based detectors with small kernels, our detector achieves a 0.7% mAP improvement over the previous best method [61], demonstrating its ability to capture long-range dependencies. For mid-range detection on Waymo Open, Table 2 compares our method to state-of-the-art methods on the validation set, where '-' indicates unavailable results. Our detector outperforms both transformer-based and large kernel-based methods, achieving gains of 1.7% and 3.3% mAPH on Level-2 difficulty compared to the transformer-based detectors DSVT [49] and PTv3 [54], respectively, while also achieving comparable results to the best small kernel-based detector [61]. For the short-range detection dataset nuScenes,Table 3 shows that our method achieves SoTA results on the mAP metric and ranks second on the NDS metric, outperforming both large kernel and transformer-based methods. The effective receptive field (ERF) is 3.3 meters for Argoverse2 and 2 meters for Waymo Open and nuScenes.

Qualitative results. To qualitatively validate our detector's ability to capture an extensive ERF, for each ground-truth object, we calculate the ERF by sampling a query voxel and analyzing its reach. Fig. 5 visualizes the ERF (color-coded from blue to red) of our detector with and without our SFM module. The more yellowish the color, the larger the ERF. For each ground-truth object, we randomly select one voxel to serve as the query and calculate the ERF associated with it. Indeed, the dependency coverage leads to large regional receptive fields. Without the SFM module (b), the receptive field is limited to local areas around each query, capturing only fragments of the objects. For instance, in the left zoom-in in Fig. 5(b), there are four vehicles. The receptive field covers only portions of each vehicle, leaving most voxels outside this field (blue). In contrast, with our SFM module (c), the detector gains a broader receptive field, allowing it to capture even partially occluded vehicles. This is reflected in the four vehicles being highlighted in green, yellow, and red colors.

Method	mAP	Vehicle	Bus	Pedestrian	Stop Sign	Box Truck	Bollard	C-Barrel	Motorcyclist	MPC-Sign	Motorcycle	Bicycle	A-Bus	School Bus	Truck Cab	C-Cone	V-Trailer	Sign	Large Vehicle	Stroller	Bicyclist	Truck	MBT	Dog	Wheelchair	W-Device	W-Ride
CenterPoint	22.0	67.6	38.9	46.5	16.9	37.4	40.1	32.2	28.6	27.4	33.4	24.5	8.7	25.8	22.6	29.5	22.4	6.3	3.9	0.5	20.1	22.1	0.0	3.9	0.5	10.9	4.2
VoxelNeXt	30.7	72.7	38.8	63.2	40.2	40.1	53.9	64.9	44.7	39.4	42.4	40.6	20.1	25.2	19.9	44.9	20.9	14.9	6.8	15.7	32.4	16.9	0.0	14.4	0.1	17.4	6.6
HEDNet	37.1	78.2	47.7	67.6	46.4	45.9	56.9	67.0	48.7	46.5	58.2	47.5	23.3	40.9	27.5	46.8	27.9	20.6	6.9	27.2	38.7	21.6	0.0	30.7	9.5	28.5	8.7
SAFDNet	39.7	78.5	49.4	70.7	51.5	44.7	65.7	72.3	54.3	49.7	60.8	50.0	31.3	44.9	24.7	55.4	31.4	22.1	7.1	31.1	42.7	23.6	0.0	26.1	1.4	30.2	11.5
SFMNet	**40.4**	79.0	49.5	72.3	49.3	46.1	65.1	73.8	55.1	53.2	62.8	54.2	30.2	44.8	26.7	57.9	31.7	20.9	8.3	33.4	44.4	21.6	0.0	22.1	3.1	33.1	12.1

Table 1. **Results on the Argoverse2 validation set (long-range, up to ± 200 meters).** The long-range detection dataset highlights our detector's ability to capture long-range dependencies.

Design	Method	mAP/mAPH L1	mAP/mAPH L2	Vehicle AP/APH L1	Vehicle AP/APH L2	Pedestrian AP/APH L1	Pedestrian AP/APH L2	Cyclist AP/APH L1	Cyclist AP/APH L2
Transformers	SST [14]	74.5/71.0	67.8/64.6	74.2/73.8	65.5/65.1	78.7/69.6	70.0/61.7	70.7/69.6	68.0/66.9
	TransFusion-L [1]	-/-	-/64.9	-/-	-/65.1	-/-	-/63.7	-/-	-/65.9
	SWFormer [45]	-/-	-/-	77.8/77.3	69.2/68.8	80.9/72.7	72.5/64.9	-/-	-/-
	CenterFormer [68]	75.6/73.2	71.4/69.1	75.0/74.4	69.9/69.4	78.0/72.4	73.1/67.7	73.8/72.7	71.3/70.2
	DSVT-Voxel [49]	80.3/78.2	74.0/72.1	79.7/79.3	71.4/71.0	83.7/78.9	76.1/71.5	77.5/76.5	74.6/73.7
	PTv3 [54]	-/-	73.0/70.5	-/-	71.2/70.8	-/-	76.3/70.4	-/-	71.5/70.4
Large kernels	LargeKernel3D [6]	-/-	-/-	78.1/77.6	69.8/69.4	-/-	-/-	-/-	-/-
	LinK [28] (6 epochs)	-/-	-/64.6	-/-	-/65.0	-/-	-/60.4	-/-	-/68.4
Small kernels	SECOND [55]	67.2/63.1	61.0/57.2	72.3/71.7	63.9/63.3	68.7/58.2	60.7/51.3	60.6/59.3	58.3/57.0
	PointPillars [23]	69.0/63.5	62.8/57.8	72.1/71.5	63.6/63.1	70.6/56.7	62.8/50.3	64.4/62.3	61.9/59.9
	Part-A2 [38]	73.6/70.3	66.9/63.8	77.1/76.5	68.5/68.0	75.2/66.9	66.2/58.6	68.6/67.4	66.1/64.9
	PV-RCNN [37]	76.2/73.6	69.6/67.2	78.0/77.5	69.4/69.0	79.2/73.0	70.4/64.7	71.5/70.3	69.0/67.8
	CenterPoint [59]	75.9/73.5	69.8/67.6	76.6/76.0	68.9/68.4	79.0/73.4	71.0/65.8	72.1/71.0	69.5/68.5
	AFDetV2 [22]	77.2/74.8	71.0/68.8	77.6/77.1	69.7/69.2	80.2/74.6	72.2/67.0	73.7/72.7	71.0/70.1
	FSD [15]	79.6/77.4	72.9/70.8	79.2/78.8	70.5/70.1	82.6/77.3	73.9/69.1	77.1/76.0	74.4/73.3
	PV-RCNN++ [39]	78.1/75.9	71.7/69.5	79.3/78.8	70.6/70.2	81.3/76.3	73.2/68.0	73.7/72.7	71.2/70.2
	VoxelNeXt [7]	78.6/76.3	72.2/70.1	78.2/77.7	69.9/69.4	81.5/76.3	73.5/68.6	76.1/74.9	73.3/72.2
	HEDNet [62]	81.4/79.4	75.3/73.4	81.1/80.6	73.2/72.7	84.4/80.0	76.8/72.6	78.7/77.7	75.8/74.9
	SAFDNet [61]	**81.8/79.9**	**75.7/73.9**	80.6/80.1	72.7/72.3	84.7/80.4	77.3/73.1	80.0/79.0	77.2.76.2
	SFMNet (ours)	**81.8**/79.8	**75.7**/73.8	80.5/80.0	72.6/72.2	85.0/80.6	77.4/73.2	79.9/78.9	77.0/76.1

Table 2. **Results on the Waymo Open validation set (mid-range, up to ± 75 meters).** Our SFMNet outperforms transformer- and large-kernel-based detectors, while achieving results comparable to the state-of-the-art small-kernel detector [61].

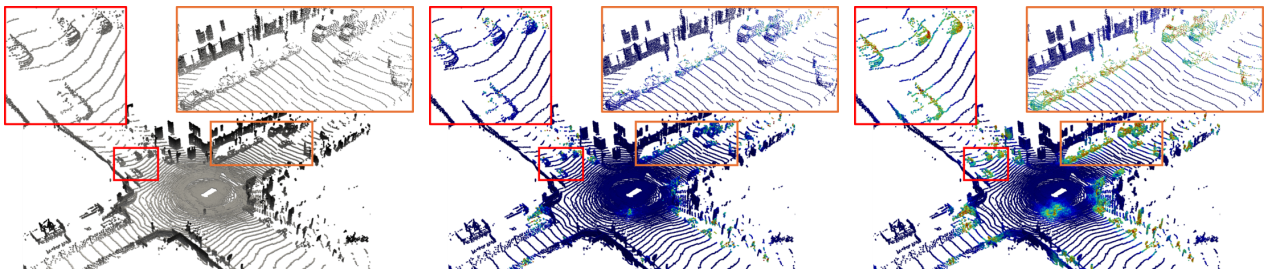
Design	Method	NDS	mAP	Car	Truck	Bus	Trailer	C.V.	Pedestrian	Motor	Bike	Cone	Barrier
Transformers	TransFusion-L [1]	70.1	65.5	86.9	60.8	73.1	43.4	25.2	87.5	72.9	57.3	77.2	70.3
	DSVT-Voxel [49]	71.1	66.4	87.4	62.6	75.9	42.1	25.3	88.2	74.8	58.7	77.8	70.9
Large kernels	LargeKernel3D [6]	69.1	63.9	85.1	60.1	72.6	41.4	24.3	85.6	70.8	59.2	72.3	67.7
	LinK [28]	69.5	63.6	-	-	-	-	-	-	-	-	-	-
Small kernels	CenterPoint [59]	66.5	59.2	84.9	57.4	70.7	38.1	16.9	85.1	59.0	42.0	69.8	68.3
	VoxelNeXt [7]	66.7	60.5	83.9	55.5	70.5	38.1	21.1	84.6	62.8	50.0	69.4	69.4
	PillarNeXt-L [24]	68.4	62.2	85.0	57.4	67.6	35.6	20.6	86.8	68.6	53.1	77.3	69.7
	HEDNet [62]	**71.4**	66.7	87.7	60.6	77.8	50.7	28.9	87.1	74.3	56.8	76.3	66.9
	SAFDNet [61]	71.0	66.3	87.6	60.8	78.0	43.5	26.6	87.8	75.5	58.0	75.0	69.7
	SFMNet (ours)	71.3	**66.9**	87.6	63.2	76.2	48.9	26.0	88.4	74.7	56.9	78.5	69.1

Table 3. **Results on the nuScenes validation set (short-range, up to ± 54 meters).** Our SFMNet outperforms all transformer- and large-kernel-based detectors, surpassing the previous SoTA small-kernel detector on the mAP metric and ranking second on the NDS metric.

4.2. Ablation study

Table 4 compares our SFM module to traditional global and local attention, as well as set attention [49], on the Argoverse2 dataset over six epochs. For this evaluation, we replaced an attention block in the 3D and 2D backbones with either a single SFM module or an alternative attention mechanism, demonstrating that our SFM achieves higher accuracy and efficiency. While global attention attains a high ERF, it quickly reaches memory limits (80 GB) due to the large voxel count. In contrast, local and set attention operate within a limited context window, with set attention outperforming local attention. However, our SFM module surpasses both in accuracy (mAP) and efficiency (memory usage) despite using a smaller context window.

Table 5 examines the effect of varying the number of SFM blocks and SRBs. Increasing the number of SRBs after the SFM block in Stage 4 (A) improves accuracy with



(a) Input scene (b) ERF without SFM (c) ERF with SFM

Figure 5. **Effective receptive field (ERF) of SFMNet.** (a) A typical input scene with height is shown in grayscale. (b,c) The ERF for each ground-truth object is color-coded from blue (outside ERF) to red (high-attention regions). Red and orange rectangles highlight zoomed-in regions. In the right (orange) zoom-in, four stationary cars and a van, plus two moving cars and a van, are shown. Without the SFM module (b), only a local portion of each vehicle is within the ERF. With SFM (c), the entire object is captured, providing more context and spatial understanding for accurate 3D detection.

Method	Window(m)	Memory(GB)	mAP
Global-attention	400	OOM	-
Local-attention	3.3	44.06	34.8
Set-attention ([49])	3.6	44.21	35.5
SFM (ours)	1.9	32.76	36.8
SFM (ours)	3.3	37.53	37.1

Table 4. **Comparison to attention.** While global attention covers the entire ERF, it quickly exceeds memory limits. In contrast, Local-attention and Set-attention [49] focus on voxels within a local context window. Our SFM module outperforms both in accuracy and efficiency, even with a smaller window.

	Stage 4		2D backbone		#Param.(M)	mAP
	#SRB	Multi.	#SRB	Multi.		
A	2	1	4	1	3.84	36.7
	4	1	4	1	4.29	37.2
	6	1	4	1	4.73	37.5
B	2	2	4	2	5.27	38.0
	2	4	4	2	7.14	38.6
	2	6	4	2	9.01	39.1
C	4	2	4	2	6.16	38.9

Table 5. **Effect of SFM blocks & SRBs.** Stacking SRBs after each SFM block improves accuracy (A). Increasing SFM repetitions boosts it further (B), but with more parameters. Configuration (C) offers a balanced trade-off.

a modest parameter increase. Similarly, adding more SFM block repetitions (Multipliers) with SRBs (B) boosts accuracy but at a higher parameter cost. A balanced approach, with moderate repetitions and additional SRBs (C), provides a good trade-off between accuracy and parameters.

Table 6 investigates the impact of different context levels with varying ERF. Increasing levels improves accuracy but also raises memory usage. Beyond a point, more levels do not help, as the network tends to ignore overly long-range features. Relying only on high dilation rates without local contexts (third row) yields suboptimal performance, highlighting the need for short-range dependencies. Additional ablations on SFM placement are in the supplemental.

#levels	dilation rates	ERF(m)	Memory(GB)	mAP
2	[1,3]	0.9	30.19	35.7
3	[1,3,5]	1.9	32.76	36.8
3	[1,5,9]	3.1	32.98	36.4
4	[1,3,5,7]	3.3	37.53	37.1
5	[1,3,5,7,9]	5.1	36.39	37.1

Table 6. **ERF effects.** We evaluate different ERF sizes by increasing the number of context levels, each using a sparse convolution with kernel size 3 and varying dilation. Four levels, yielding an ERF of 3.3 meters, achieve the best accuracy.

Limitations. Our method is designed to capture long-range dependencies and enhance models that lack this capability. On datasets requiring long-range reasoning, it improves not only those models but also others with a similar objective. However, on short-range data, while it still benefits the former, its performance is on par with the latter.

5. Conclusion

This paper introduced a novel 3D Focal Modulation module, SFM, designed to effectively capture dependencies from short- to long-range. It is based on a hierarchical submanifold sparse convolution design to efficiently handle data sparsity. It also presented SFMNet, a new sparse convolution-based 3D detector built on top of the SFM. It addresses the limitation of traditional sparse convolution detectors by significantly enlarging the effective receptive field. Experimental results demonstrate that SFMNet captures dependencies very well and achieves SoTA performance on multi-range datasets.

Acknowledgments. This work was supported by the Israel Science Foundation (ISF) under grant 2329/22, ADRI-Technion, Elbit, and the Israeli Smart Transportation Research Center (ISTRC).

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 7

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 6

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 1

[4] Honghao Chen, Yurong Zhang, Xiaokun Feng, Xiangxiang Chu, and Kaiqi Huang. Revealing the dark secrets of extremely large kernel convnets on robustness. *arXiv preprint arXiv:2407.08972*, 2024. 2

[5] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9775–9784, 2019. 1, 2

[6] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13488–13498, 2023. 2, 3, 4, 7

[7] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 1, 2, 7

[8] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023. 2, 3

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1

[10] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *arXiv preprint arXiv:2012.15712*, 1(2):4, 2020. 1, 2

[11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 2, 3

[12] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5513–5524, 2024. 2, 3, 5

[13] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Rangedet: In defense of range view for lidar-based 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2918–2927, 2021. 1

[14] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022. 2, 3, 7

[15] Lue Fan, Feng Wang, Naiyan Wang, and Zhao-Xiang Zhang. Fully sparse 3d object detection. *Advances in Neural Information Processing Systems*, 35:351–363, 2022. 7

[16] Tuo Feng, Wenguan Wang, Fan Ma, and Yi Yang. Lsk3dnet: Towards effective and efficient 3d perception with large sparse kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14916–14927, 2024. 2, 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[18] Ben Graham. Sparse 3d convolutional neural networks. *arXiv preprint arXiv:1505.02890*, 2015. 3

[19] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 1, 3

[20] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 1

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[22] Yihan Hu, Zhuangzhuang Ding, Runzhou Ge, Wenxin Shao, Li Huang, Kun Li, and Qiang Liu. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 969–979, 2022. 7

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar 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. 1, 2, 7

[24] Jinyu Li, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar

point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17567–17576, 2023. 7

[25] Lin Liu, Ziying Song, Qiming Xia, Feiyang Jia, Caiyan Jia, Lei Yang, Yan Gong, and Hongyu Pan. Sparsedet: a simple and effective framework for fully sparse lidar-based 3d object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[26] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022. 2, 3

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[28] Tao Lu, Xiang Ding, Haisong Liu, Gangshan Wu, and Limin Wang. Link: Linear kernel for lidar-based 3d perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1105–1115, 2023. 2, 3, 7

[29] Xu Ma, Xiyang Dai, Jianwei Yang, Bin Xiao, Yinpeng Chen, Yun Fu, and Lu Yuan. Efficient modulation for vision networks. *arXiv preprint arXiv:2403.19963*, 2024. 4

[30] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 1

[31] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3164–3173, 2021. 3

[32] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7463–7472, 2021. 3

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[35] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[36] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[37] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 1, 2, 7

[38] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020. 1, 2, 7

[39] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023. 1, 2, 7

[40] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1711–1719, 2020. 1, 2

[41] Oren Shrout, Yizhak Ben-Shabat, and Ayellet Tal. Gravos: Voxel selection for 3d point-cloud detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21684–21693, 2023. 1, 2

[42] Ziying Song, Haiyue Wei, Caiyan Jia, Yongchao Xia, Xiaokun Li, and Chao Zhang. Vp-net: Voxels as points for 3-d object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 1

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 2, 5, 6

[44] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2021. 1

[45] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *European Conference on Computer Vision*, pages 426–442. Springer, 2022. 2, 3, 7

[46] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, et al. Stanley: The robot that won the darpa grand challenge. *Journal of field Robotics*, 23(9):661–692, 2006. 1

[47] Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in Neural Information Processing Systems*, 35:34899–34911, 2022. 1

[48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[49] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. Dsvt: Dy-

namic sparse voxel transformer with rotated sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13520–13529, 2023. 1, 2, 3, 6, 7, 8

[50] Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6792–6802, 2023. 2, 3

[51] Li Wang, Ziying Song, Xinyu Zhang, Chenfei Wang, Guoxin Zhang, Lei Zhu, Jun Li, and Huaping Liu. Sat-gcn: Self-attention graph convolutional network-based 3d object detection for autonomous driving. *Knowledge-Based Systems*, 259:110080, 2023. 2

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2, 3

[53] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1, 2, 4, 6

[54] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 6, 7

[55] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2, 7

[56] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 4

[57] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020. 1, 2

[58] Maosheng Ye, Shuangjie Xu, and Tongyi Cao. Hvnet: Hybrid voxel network for lidar based 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1631–1640, 2020. 1, 2

[59] Tianwei Yin, Xingyi Zhou, and Philipp 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, 2, 5, 7

[60] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 5

[61] Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li, Si Liu, and Xiaolin Hu. Safdnet: A simple and effective network for fully sparse 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14477–14486, 2024. 1, 2, 6, 7

[62] Gang Zhang, Chen Junnan, Guohuan Gao, Jianmin Li, and Xiaolin Hu. Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 7

[63] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 3

[64] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3555–3562, 2021. 1, 2

[65] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14494–14503, 2021.

[66] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1, 2

[67] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020. 5

[68] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022. 2, 3, 7

[69] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017. 1