

AttentiveGRU: Recurrent Spatio-Temporal Modeling for Advanced Radar-Based BEV Object Detection

Loveneet Saini

Aptiv

Wuppertal, Germany

sainiloveneet.ls@gmail.com

Mirko Meuter

Aptiv

Wuppertal, Germany

mirko.meuter@aptiv.com

Hasan Tercan, Tobias Meisen

University of Wuppertal

Wuppertal, Germany

{tercan,meisen}@uni-wuppertal.de

Abstract

Bird's-eye view (BEV) object detection has become important for advanced automotive 3D radar-based perception systems. However, the inherently sparse and non-deterministic nature of radar data limits the effectiveness of traditional single-frame BEV paradigms. In this paper, we address this limitation by introducing AttentiveGRU, a novel attention-based recurrent approach tailored for radar constraints, which extracts individualized spatio-temporal context for objects by dynamically identifying and fusing temporally correlated structures across present and memory states. By leveraging the consistency of object's latent representation over time, our approach exploits temporal relations to enrich feature representations for both stationary and moving objects, thereby enhancing detection performance and eliminating the need for externally providing or estimating any information about ego vehicle motion. Our experimental results on the public nuScenes dataset show a significant increase in mAP for the car category by 21% over the best radar-only submission. Further evaluations on an additional dataset demonstrate notable improvements in object detection capabilities, underscoring the applicability and effectiveness of our method.

1. Introduction

Perception systems in today's vehicles support a wide range of applications, such as automatic emergency braking (AEB) and automatic valet parking (AVP). Thereby, the standard industry approach is to use a multi-sensor suite of radar, lidar, and cameras to facilitate these complex applications. Radar, known for its superior long-range capabilities and effectiveness in adverse weather conditions, as well as its Doppler measurement feature, has become an integral part of these perception systems. The automotive 3D radar system processes reflections from targets through a signal processing chain that estimates parameters such as time-

of-flight and arrival angle, and produces range-azimuth-Doppler (RAD) feature maps of the scene within the radar's field of view. Despite its lower resolution and sparser data compared to lidar and cameras, radar's ability to capture point reflections from objects provides unique object detection capabilities through distinct target signatures in the radar measurement space.

Recent deep learning approaches to radar data have explored a variety of input formats. The state of the art includes models that process a single radar frame either in a point cloud format, emphasizing specific local features [19], or in a Bird's Eye View (BEV) grid format, facilitating the natural extraction of local patterns [17, 32]. Notably, due to emerging trend of fusing different sensors like camera and radar in common BEV space, the research in radar exclusive BEV based detection has become of significant importance as it helps in establishing stronger backbones in such multi-modal architectures.

Radar, as a low-resolution sensor, is inherently affected by sparsity issues [33, 34]. To mitigate this, some single-frame approaches have employed the concatenation of multiple radar frames to generate a denser input [23]. However, this method of aggregation is computationally demanding and fails to leverage the advantages of exploiting temporal relationships. Hence, in this work, we propose a model that harnesses the temporal relationship patterns in BEV radar data using recurrent methods.

For denser input modalities like lidar, recent research underscores the advantages of incorporating temporal relationships. These multi-frame approaches [5, 8] generally begin by aligning pixel-wise BEV features from historical data with current inputs to compensate for ego-vehicle motion, utilizing either a learned or externally provided ego-pose matrix, and subsequently integrate them using ConvGRU [1] or attention mechanisms [27]. A similar and more up-to-date approach is utilized for camera images, where historic information is pre-aligned using ego-vehicle information [9, 14]. Specifically in [14], Li et al. have noted that the availability of an ego pose matrix allows the align-

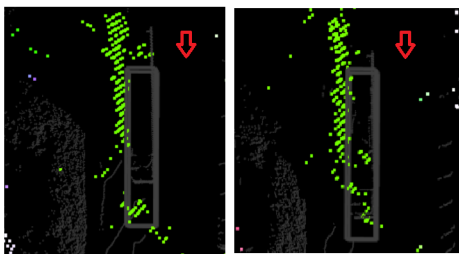


Figure 1. Radar reflections for two consecutive scans from a truck approaching towards ego vehicle in the direction of arrow

ment of only stationary objects due to the varying dynamics of moving objects in the scene. Following such pre-alignment with external ego information, Li et al. have proposed using deformable attention to model temporal relations by concatenating features from past and present 2D BEV queries. However, these lidar and camera specialized methods are less robust to the constraints posed by commercial radar data (demonstrated below) unless high-resolution radar imaging techniques are employed, which comes with increased pre-processing, data requirements, and costs.

In the radar measurement space, objects are detected by capturing point reflections of radar signal from objects (see Figure 1). Due to longer wavelengths of the radar sensor [33, 34] the representations of objects in this space can be highly sparse with high dynamic noise. Specifically, the highly sparse reflections from the same object across consecutive scanned frames may vary in terms of the part of the object they originate from (front of truck in Figure 1), rendering them non-deterministic. Furthermore, unlike camera and lidar, the radar receiver antenna exhibits the highest sensitivity at the center of its beam (boresight), with sensitivity diminishing at greater angles. This attribute leads to progressively weaker observability of objects as they move away from the boresight. Consequently, depending on the dynamics of the object, its representation may appear smeared and significantly noisy when observed across consecutive frames.

The above mentioned challenges associated with the radar feature maps makes the direct use of temporal fusion methods from lidar and camera non-trivial. However, radar’s ability to provide depth (range) and velocity (Doppler) of targets provides a unique opportunity to correlate objects over time. In typical driving scenarios, where objects do not abruptly change their motion, the range-azimuth features facilitate object matching across spatial depths, while the velocity feature aids in tracking moving objects across adjacent scans. In this paper, we present a novel plug and play temporal fusion layer for improving existing radar based object detectors for the widely used Range-Azimuth-Doppler (RAD) radar input format. Our proposed layer has linear computational complexity and is

designed to improve object detection performance without the need for providing or estimating ego motion based transformations. Specifically, our model incorporates a module inspired by the Perceiver [10], which exploits the consistency of an object’s latent encodings (query) across time frames to identify and fuse correlated object structures in a recurrent manner. We benchmark our proposed model using the public nuScenes dataset [4] and additionally use a larger and more complex proprietary radar dataset from [3] for rigorous evaluation and detailed ablation studies. In summary, the contributions of this paper are as follows:

- We propose a temporal fusion layer tailored for radar data, to enhance detection performance by integrating individualized spatio-temporal contexts for each object recurrently over input time frames. This layer significantly reduces the computational effort by having linear complexity in both space and time and brings advantages of temporal fusion to highly sparse and challenging inputs.
- We address a gap in the existing radar literature by combining the recurrence of GRUs with the attention mechanism to introduce a novel ”attention gating” approach. This enhancement improves the ability of the GRU to selectively focus on temporally correlated locations for both moving and stationary objects across time frames, without the need for externally provided ego-motion information or explicit motion estimation.
- We present an ablation study to analyze how global receptive fields and spatio-temporal representations influence detection performance for different object types, sizes and dynamics. This evaluation integrates the proposed layer with both traditional Feature Pyramid Networks (FPN) and modern transformer-based BEV object detection models.

2. Related Work

Multi-Frame object detection: Only a handful of recent works have addressed radar-only object detection in multi-frame scenarios. The studies by [13] and [30] model temporal relations by employing masked self-attention mechanisms between potential object centerpoints (referred as top K object features) from consecutive time frame. As noted in their research, this method has limitations when applied to data from the nuScenes dataset, which uses commercial radar sensors, and is more appropriate for high-resolution radar images from advanced radar sensors [21]. To overcome such limitations, we propose to perform temporal fusion using the latent space [10] of objects, instead of fusing over keypoints. By leveraging the consistency of an object’s latent representation (query) across time frames, our method facilitates direct fusion over pixel-wise BEV features and brings the advantages of temporal fusion to highly sparse and challenging inputs. This is supported by observations that temporal fusion over pixel-wise BEV features

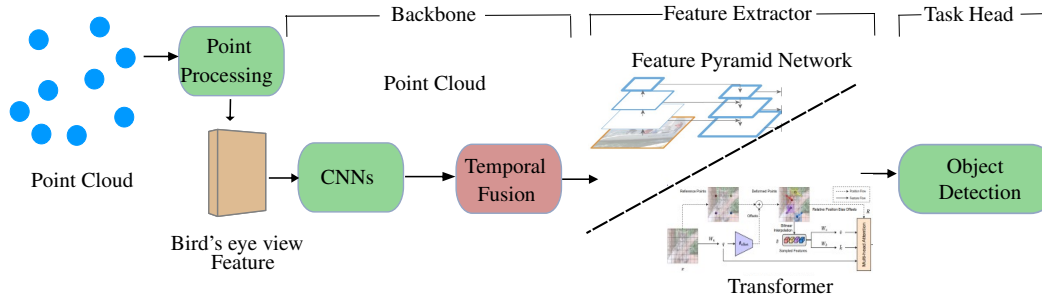


Figure 2. Overall architecture with temporal fusion layer in the backbone

is more beneficial for improving performance on sparser inputs than fusion over the keypoints.

A recent extension of [13] and [30] by Pu Wang et al. in [31] explore temporal relations over a sequence of time steps rather than just consecutive frames. Their method models temporal relations using attention applied to potential object center points extracted from past T time steps. Specifically, they sample K subset features (extracted as K center points with heatmaps) from each step's feature map using the same approach as in [30] and group them into window patches. By exploiting the temporal order of the input sequence (termed "deformable temporal order"), these patches are reorganized into a new window that collects subset features from all past T time steps in temporal order. A masked self-attention mechanism is then used to update each subset feature. However, as the authors note, their approach is suitable for high-resolution radar image datasets and has limitations when applied to sparse nuScenes data [13]. Furthermore, it has quadratic complexity with respect to the sequence length. In contrast, our approach directly fuses sampled BEV pixels instead of center points, allowing it to handle sparser inputs from commercial sensors more effectively than [31]. In addition, by integrating recurrence with an attention mechanism, our method maintains linear computational complexity in both space and time. In contrast to fusion over keypoints, authors in [7] have proposed the use of Convolutional LSTM layer[22] for temporal fusion by performing temporal fusion over different 2D views of 3D range-azimuth-Doppler (RAD) measurement space. However, as noted in [5], ConvLSTM is primarily effective only for stationary objects, since ego compensation is only able to align them across frames, by compensating for the movement of the ego vehicle but cannot align moving objects (each with different dynamics).

In contrast to these radar-based approaches, our method eliminates the need to pre-align objects prior to temporal fusion, thereby avoiding explicit motion estimation effort for moving objects and even bypassing reliance on externally providing ego-motion information (using extra sensor) for stationary objects. Specifically, our approach dynamically

identifies and samples temporally correlated locations for each object separately in both present and memory states. These locations are then recurrently fused via a state integration module, as described in Section 4.3.2. The learnable, deformable nature of this sampling mechanism across space and time allows the model to effectively adapt to the motion of both individual objects and the ego vehicle over time.

3. Architecture Overview

Our work extends the existing CNN and transformer based models for radar BEV object detection into the spatio-temporal domain to exploit temporal relationships from the RAD feature map of the scene within the radar's field of view. Traditionally, FPNs have been long used in many radar object detectors over the years [17, 26, 32], while only recently transformer based detectors have emerged [20]. In order to utilize temporal relation, we first design a flexible architecture template which has a replaceable feature extractor to allow for substituting either FPN or a transformer. To this end, we adopt the popular Fully Convolutional One-Stage (FCOS) object detection framework [25] as the basic template for our radar input as shown in Figure 2. We adapt the original FCOS architecture such that it uses an input of pre-processed radar reflections by incorporating (i) a grid processing backbone with a temporal fusion module, (ii) a replaceable feature extractor, and (iii) a center point object detection head.

3.1. Radar Signal Point-Processing

Figure 3 illustrates the standard radar signal processing chain used to process the received reflections at each point. First, the radar signal undergoes a 2D Fast Fourier Transform (FFT), which facilitates the generation of a range Doppler spectrum. This spectrum is then processed through a Constant False Alarm Rate (CFAR) detector designed to effectively extract radar targets. Once detected, the arrival directions (angles) of the targets are estimated. This process is applied to all incoming radar point clouds to obtain common point features. The processed point cloud is then pro-

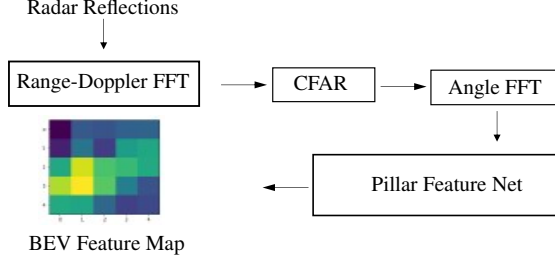


Figure 3. Per-point processing chain for input radar reflections. Blue regions in the BEV map represent sparsity, while yellow is the peaks from reflections

jected onto a 2D BEV grid, transforming it into an image-like input. For this projection, we use the Pillar Feature Net from the PointPillars methodology [12], which divides the cloud space into pillars similar to vertical bins. This approach is prevalent in much of the existing literature, such as [11, 24, 26, 28], and is therefore adopted here “as is”.

3.2. Backbone

In the backbone stage, the input BEV grid is first processed through convolution layers to introduce spatial awareness of the grid input into the network (see Figure 2). The output sequence from these convolution layers is then downsampled by a factor of K using a dynamic convolution module [6]. In order to improve the representational capacity for low-resolution objects in the input, this module has three parallel kernels $K = 1, 2, 4$. A crucial component of this stage is the Temporal Fusion with AttentiveGRU module, which is responsible for extracting temporal relations. We will discuss this module in more detail in section 4.

3.3. Feature Extractor

We designed our architecture to be flexible, allowing for a swappable feature extractor that can utilize either an FPN network or a transformer. Figure 2 illustrates the FPN-based feature extraction stage [15], where the input is processed by multi-level CNNs to extract features at different resolutions, incorporating max-pooling and up-sampling layers. Additionally, as described, the FPN can be replaced by a transformer, as demonstrated in [20], enabling the use of a global receptive field that effectively captures global object features of varying sizes.

3.4. Head

For the detection head, the target processor is inspired by the FCOS [25] network, which uses a centerness score for keypoint estimation with heat map regression to identify and collect top-K center candidates. As previously demonstrated in Tian et al. research, FCOS is a suited alternative to anchor-based methods due to its capacity to utilise a multitude of foreground samples for the purpose of training the

regressor. This makes it an optimal candidate for the detection head. The identified center candidates are finally regressed to obtain object properties after training using a focal loss [16], similar to FCOS.

4. Temporal Fusion with AttentiveGRU

4.1. Overview

Figure 4 shows our proposed temporal fusion block, which consists of an attention gating and a state integration module. As shown in the figure, the block has three inputs, namely memory state (single BEV register buffer), present state (single BEV frame), and M learnable latent queries (parameters). For an input sequence of length T , to the overall network (Figure 2), each input frame is fed in a recurrent manner, starting from time step $t=0$ to $t=T$, as the current state, in the Figure 4. Thus, the entire input sequence passes through the fusion block in T passes. At each pass, except the last, the fusion output is used as an updated memory block, while the final T^{th} pass fusion output is fed to the detector head. The goals of the block are first to dynamically learn temporally correlated locations across present and memory states for each object in the input using attention gating, and later to sample features from these locations and fuse them using the state integration module.

4.2. Mechanism

To illustrate the working principles of our fusion block, we detail its operation and theoretical underpinnings in terms of its individual modules as shown in Figure 4.

4.2.1. Attention Gating

Each of the M different learnable queries learn a latent-encoding for M different objects. For each object, the mechanism of detecting correlated locations over time is based on the idea of existence of a unique latent encoding that remains consistent across time frames. This idea has its theoretical motivation in the work Perceiver [10], in which Jaegle et al. proposed that an input modality (byte array) can be encoded into a latent representation (i.e., latent query), where each latent query serves as the centroid of clusters present in the input. Using visualization of attention maps, Jaegle et al. have extensively shown the feasibility of such encoding for diverse input images across the datasets and even other data modalities.

We extend this concept across multiple input time frames and propose that the latent representation query for each object (cluster) in the inputs (M latent queries for M different objects) is distinct and remains stable across time scans. Therefore, given an input of historical and present states, we can decode a learnable latent query to dynamically learn correlated pair of locations for each object over time, and then sample them to extract temporal features. Notably, the core assumption of our method will not hold

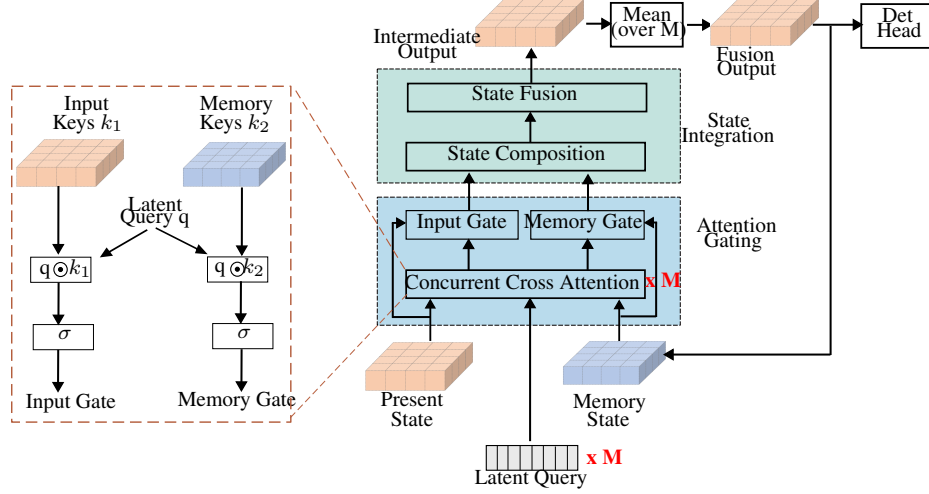


Figure 4. Illustration of the proposed Temporal Fusion Block with Attention Gating (AG) and State Integration (SI) modules

in sudden acceleration and braking scenarios where conditions are rapidly changed. While our method is effective for typical driving scenarios, it may not generalize to out-of-distribution cases with such dynamic scenarios.

Dynamically learning correlated pair of locations across time frames effectively addresses the non-deterministic and noisy nature of radar inputs, as the latent representation of the object remains consistent even when representations are smeared across adjacent frames.

As shown in the Figure 4, M latent queries are used to dynamically learn locations over both historical memory state and present state for M different objects. Specifically, a novel Concurrent Cross Attention operation is used where each of the query q jointly attends the present (keys k_1) and memory (keys k_2). On the resulting attention maps, a gating function σ is applied to map attention scores to binary gates for dynamic selection of locations from present and memory. To make the dynamic selection process adaptive to the input, the gating function σ is modeled as a learnable thresholding of the sigmoid layer. Essentially, a sigmoid layer processes the attention scores and the median of the sigmoid layer output is used as a learnable threshold resulting in binary gates. For radar data, the learned features of each of the latent query are higher dimensional embeddings of the RAD input to the overall network, so the dynamically learned locations are partially revealed using the spatial depth (range) and (Doppler) velocity information across frames.

The output of attention gating results in $2 \times M$ binary gates corresponding to each of the M latent queries for the memory and present states, respectively. As the same latent query is used for both states, this facilitates the revelation of correlated locations through the corresponding gate. Consequently, at the input of the state integration module, each

gate controls what information is passed on to the module by effectively populating the locations where the gate value is 1 with features from the corresponding state. This operation results in a correlated pair of gated states from the memory ($h_{t-1}(q)$) and present state ($x_t(q)$) for each latent query q .

4.2.2. State Integration

To extract temporal features from the resulting pairs of gated states, we modify the processing of vanilla ConvGRU [1]. The vanilla version of the operator directly process the entire input by using reset and updates gates as modeled by equations 1-4 in the paper [1]. Since, we have gated states for each object separately available, rather than the entire input, we reformulate the original equations to instead process our gated states ($h_{t-1}(q)$ and $x_t(q)$) for temporal features extraction. Specifically, for each pair of gated states we create a composite state $C(q)$ by concatenating gated states ($[h_{t-1}(q), x_t(q)]$) for each of the M queries q (called the state composition in Figure 4). Such composite state contains overlapping and non-overlapping regions of the same object across time in feature space, as shown in the Figure 6. Hence, we can combine these areas to extract BEV temporal features for each object separately. For overlapping areas, we use local 1×1 convolution based gates $l_1(q)$ and $l_2(q)$ to obtain the fused state $h_t(q)$ as explained in equations 1. In contrast to the global gating operation of attention gating module, the use of local gate facilitates sequential processing for each pixel in present and memory state and allows the extraction of temporal features that detail the object's local movement across its own spatial dimensions (left of Figure 6). Based on overlapping features, gate $l_1(q)$ is used to control information flow from $h_{t-1}(q)$ to generate updated composite state $\tilde{C}(q)$. This generated state is again

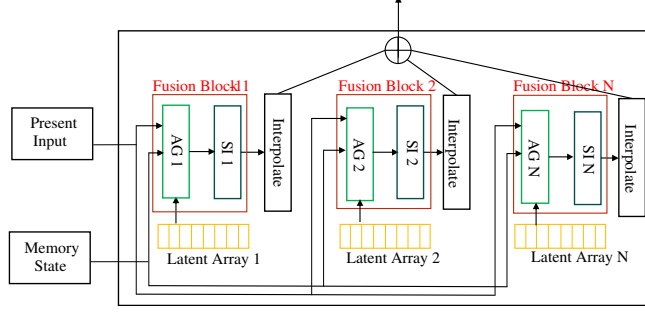


Figure 5. An exemplary AttentiveGRU layer, showcasing the use of proposed Temporal fusion blocks with Attention gating (AG) and State Integration (SI) modules

gated by gate $l_2(q)$ which controls information flow from present with gated past state. Finally, the complement gate $(1 - l_2(q))$ restores the past gated information.

$$\begin{aligned}
 C(q) &= [h_{t-1}(q), x_t(q)] \\
 l_1(q) &= \sigma(\text{Conv1d}(C(q))), \\
 l_2(q) &= \sigma(\text{Conv1d}(C(q))), \\
 \tilde{C}(q) &= \tanh(\text{Conv1d}([l_1(q) * h_{t-1}(q), x_t(q)])) \\
 h_t(q) &= (1 - l_2(q)) * h_{t-1}(q) + l_2(q) * \tilde{C}(q) \\
 h_t(q) &= \text{DeformConv}(h_t(q))
 \end{aligned} \tag{1}$$

Following this, we process the non-overlapping regions by use of a deformable convolution [35] using a kernel of size S . Such an operation facilitates convolution at offset-defined locations (indicated by green arrow lines in right of Figure 6). The regressed deformable offsets for each location, bridges the gaps between non-overlapping regions and facilitate the convolution and thus extract temporal features related to the global movement of the object.

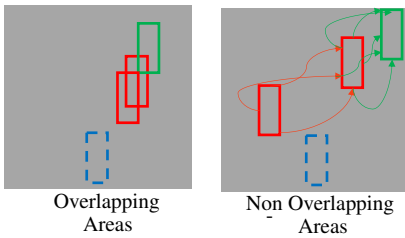


Figure 6. Visualization of composite state $C(q)$: Blue indicates the ego vehicle, red marks gated state from the memory ($h_{t-1}(q)$), and green represents the one from present ($x_t(q)$)

The operations in equations 1 (State Fusion in Figure 4) are performed for each of M latent queries to selectively obtain intermediate output for each object based on pixel-wise BEV features. The Fusion Output is generated by averaging all temporally fused gated states for each latent query.

4.3. Block Aggregation

With the above procedure, all latent queries should ideally fuse information about every object in the present state. However, this is a strong assumption, since not all latent queries may capture all relevant objects, and even individual latent queries may not correspond to a complete object, but rather to specific facets of the same object. In order to address this, it is necessary to utilise more than one fusion block in order to ensure sufficient flexibility enabling the capture of both complete objects and all objects over time. Figure 5 illustrates the arrangement of the fusion layer for N fusion blocks. Multiple fusion blocks (with shared memory) are used to ensure that all objects are captured from the sequence of input frames and memory over time. The outputs of each fusion block are interpolated and aggregated to produce the final generalized feature representation. These connections are designed to allow the gradient to flow directly to each of the M latent queries of each fusion block. Since no auxiliary losses are used, this is an important step to ensure that the gradient is not diminished for all learnable latent queries, as it would be in a feed-forward style.

5. Experiments

5.1. Dataset

Given the large domain gap in terms of sensor characteristics between different radar sensors and levels of radar data, we tested our model on two different datasets. The nuScenes dataset restricts the number of submissions for its test dataset, prompting us to perform our comprehensive comparison with all relevant works, reimplemented from the literature, on its official validation dataset, following the test approach in [26]. Furthermore, although our network is capable of detecting object types such as bikes and pedestrians, it is important to note that the current state-of-the-art for radar-only detectors on the nuScenes dataset predominantly focuses on evaluating models based on their performance in detecting the 'car' class. This has become a de facto standard evaluation approach, as reflected in the works

Table 1. Quantitative benchmark for class *car* on the nuScenes dataset

Object detector	Type	AP4.0(%) $\uparrow$	mAP (%) $\uparrow$	rel. mAP $\uparrow$
GNN [23]	point-based	24.7	13.7	-57.7%
PointPillars [12]	grid-based	37.0	22.0	-32%
RPFA-Net [28]	grid-based	38.3	23.1	-28.7%
KPConvPillars [26]	hybrid	42.2	26.2	-19%
CenterPoint Tf. [20]	grid-based	43.4	32.4	baseline
AttentiveGRU (ours)	grid-based	46.7	36.9	+14%
AttentiveGRU (ours) + CenterPoint Tf.	grid-based	50.4	39.2	+21%

[18, 20, 23, 26, 29]. Consequently, in line with this common practice in literature, our quantitative evaluation on the nuScenes dataset focuses primarily on the "car" class, which accounts for the majority of the bounding boxes in the dataset. In our work, we use standard validation and training split of public nuScene dataset.

In order to conduct a comprehensive ablation of our model, we utilize an additional proprietary radar dataset from [3] having a higher complexity. This dataset features finer class separations (see Table 2) based on vehicle dynamics (moving and stationary), and vehicle sizes (regular to large) enabling a more rigorous evaluation of our model. Large vehicles include types such as trucks and buses, while regular vehicles include cars, auto-rickshaws, etc. Furthermore, this dataset includes challenging real-world scenarios, such as fully occluded objects that are invisible to both lidar and camera, but detectable in radar data through multipath propagation. In our work, for this additional dataset we use a total of 21766 scenes for training and 9294 scenes for the test set.

Table 2. Class Categories for additional radar dataset

Classes
Vehicle Moving (VM)
Vehicle Stationary (VS)
Large Vehicle Moving (LVM)
Large Vehicle Stationary (LVS)
Bikes
Pedestrians

5.2. Implementation Details

We trained all the networks presented here using the Adam optimizer with a learning rate of 1×10^{-4} for twenty-nine epochs and a rate of 1×10^{-5} for the last 30th epoch using early stopping, with a batch size of 1, and on an Nvidia 2080TI GPU. The models are fitted using the sigmoid focal-based loss [16] to account for class imbalance in the training dataset. For the bounding box regression, the smooth L1 loss is used. For the FPN feature extractor, a 3-level

pyramid with a feature dimension size of 64 is employed throughout the architecture, while the transformer block utilizes the default parameter values from [20]. In the proposed fusion layer, the latent query dimension for each fusion block is set to 64 with a total of $M = 32$ latent queries. A kernel size $S = 3$ is selected for deformable convolution in state integration module and input of time sequence length $T=2$ is used for nuScene dataset while $T=12$ is used for proprietary dataset. In Figure 4, the mean is taken before applying deformable convolution to the intermediate output to speed up computations for the nuScenes dataset, as it is highly sparse. In total, the fusion layer contains three fusion blocks, resulting in a total of $32 \times 3 = 96$ latent queries.

5.3. Results

For performance comparison, we adhered to the evaluation strategies employed by prior state-of-the-art works in automotive radar research, utilizing the nuScenes dataset. In the presented comparison of the proposed model with other detectors, we use the quantitative analysis presented in [20, 26], which reimplements and evaluates models from the literature on the nuScenes dataset with radar data as exclusive input. We choose the current best approach of Centerpoint Transformer [20] (for nuScenes in 2024 with training and inference using radar only modality) as our baseline. Additionally, as described in Section 2, we do not compare our model to the related works by Pu Wang et al. [13, 30, 31] because they are designed for high-resolution radar images from advanced sensors and as noted in [31], are unsuitable for the sparse data of the nuScenes data [13].

Table 1 shows the results of our experiments on nuScenes. Comparing the mean average precision (average: AP4.0, AP2.0, AP1.0 and AP0.5), it can be observed that compared to the current state-of-the-art transformer based model [20] 13.8% improvement can be seen by integrating temporal relations in CNN based FPNs. This impact is further explored in the following ablation studies. Notably, the best performance is obtained by substituting our proposed AttentiveGRU layer directly into the transformer [20] for temporal fusion, which brings additional 21% improvement

over the previous state-of-the-art approach.

6. Ablation

6.1. Setup

Unlike images, the non-deterministic (Figure 1) and sparse (Figure 3) nature of radar data, particularly with BEV images in polar coordinates (range-azimuth), makes it challenging to effectively illustrate the added value using only a few selected images. Following the common practice in radar literature [18, 23], we provide a rigorous ablation of our methods, beyond scalar metrics, by analyzing detailed precision-recall curves for different class categories. To this end, we perform the subsequent ablation studies on the proprietary dataset [3], which allows for rigorous evaluation of the models by featuring finer class separations. Furthermore, with such class categories, we analyze how concepts of global receptive fields and spatio-temporal fields influence the final detection performance for objects based on their type, size and dynamics.

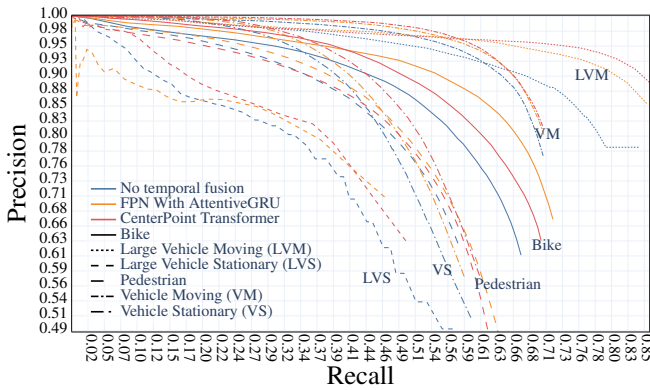


Figure 7. Precision-Recall curves illustrating the detailed performance of the proposed AttentiveGRU block

6.2. Performance Curves

Figure 7 compares the performance for the object detection task across all precision and recall metrics for bikes, pedestrians along-with large and regular classes of moving and stationary vehicles. Two of the curves in this figure shows performance of our FPN based model with and without AttentiveGRU module. Furthermore, a third curve shows performance of the current state-of-the-art model taken directly from [20] without integrating it with the AttentiveGRU layer.

6.3. Discussion

It can be observed that the addition of temporal fusion for the FPN based model brings a significant performance improvement for all classes under all operating conditions, indicated by a larger area under the curve (AUC). This im-

provement means that any chosen operating point will provide superior performance compared to alternatives, without requiring a trade-off between precision and recall metrics. In particular, the addition of AttentiveGRU results in a significant performance gain for the moving classes such as Bike, which have micro-Doppler signal [2], resulting in a unique latent query representation and better identification and fusion of gated states.

One of the advantage of using performance curves is that it can help to outline the role of global receptive field versus the temporal field for detecting objects of different class categories under different recall-precision tradeoffs. Considering the AUC metric for the curves of the CenterPoint transformer and FPN with AttentiveGRU, it can be observed that, across all operating conditions of the models, temporal fusion outperforms the (higher) global receptive field of transformers for regular-sized objects such as the 'bike' and 'pedestrian' classes. In contrast, bigger objects from LVM and LVS classes favors larger receptive field, while regular moving vehicle (VM) class benefits equally from both. For stationary vehicles, due to absence of dynamics, global receptive field is more favorable. In summary, the ablation studies highlight that temporal and global receptive fields offer distinct advantages for different object classes. Therefore, it is understandable that the model combining transformers with temporal fusion using AttentiveGRU delivers the best performance, as demonstrated in Table 1.

7. Summary

In this paper, we have shown that the detection performance of objects of typical type, size and motion benefits from the introduction of spatio-temporal receptive fields in object detectors. We have presented an AttentiveGRU layer tailored to radar constraints, with an efficient attention gating module that dynamically identifies correlated object structures (pixel-wise BEV features) across radar scans for each object in the scene. These identified structures are fused with a state integration module to extract temporal features by processing overlapping and non-overlapping regions over time for each object. To make the fusion efficient while still being able to process challenging sparse inputs, we use recurrence together with an attention mechanism to fuse directly over sampled BEV pixel features. As a result, our approach, with linear complexity in space and time, demonstrates a significant performance improvement over existing radar-only object detection models for the challenging nuScenes dataset and an additional dataset, without reliance on external ego information or estimation effort. Furthermore, the modular design of the layer allows it to be easily incorporated into many radar or multimodal detection models. In our future work, we intend to enhance the multitasking capabilities of the architecture by incorporating complex tasks such as scene flow to improve the exploitation of correlated object structures beyond temporal feature extraction.

References

- [1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations, 2016. [1](#), [5](#)
- [2] Domenic Belgiovane and Chi-Chih Chen. Micro-doppler characteristics of pedestrians and bicycles for automotive radar sensors at 77 ghz. In *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pages 2912–2916. IEEE, 2017. [8](#)
- [3] Marco Braun, Moritz Luszcz, Jan Siegemund, Kevin Kollek, and Anton Kummert. Quantification of uncertainties in deep learning-based environment perception. In *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, pages 1–8. IEEE, 2021. [2](#), [7](#), [8](#)
- [4] 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 multi-modal dataset for autonomous driving, 2020. [2](#)
- [5] Ernesto Lozano Calvo, Bernardo Taveira, Fredrik Kahl, Niklas Gustafsson, Jonathan Larsson, and Adam Tonderski. Timepillars: Temporally-recurrent 3d lidar object detection, 2023. [1](#), [3](#)
- [6] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. [4](#)
- [7] Colin Decourt, Rufin VanRullen, Didier Salle, and Thomas Oberlin. A recurrent cnn for online object detection on raw radar frames. *IEEE Transactions on Intelligent Transportation Systems*, 2024. [3](#)
- [8] Jinghua Hou, Zhe Liu, Zhikang Zou, Xiaoqing Ye, Xiang Bai, et al. Query-based temporal fusion with explicit motion for 3d object detection. *Advances in Neural Information Processing Systems*, 36, 2024. [1](#)
- [9] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection, 2022. [1](#)
- [10] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. [2](#), [4](#)
- [11] Daniel Köhler, Maurice Quach, Michael Ulrich, Frank Meinel, Bastian Bischoff, and Holger Blume. Improved multi-scale grid rendering of point clouds for radar object detection networks. In *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023. [4](#)
- [12] 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. [4](#), [7](#)
- [13] Peizhao Li, Pu Wang, Karl Berntorp, and Hongfu Liu. Exploiting temporal relations on radar perception for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17071–17080, 2022. [2](#), [3](#), [7](#)
- [14] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers, 2022. [1](#)
- [15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. [4](#)
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [4](#), [7](#)
- [17] Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [1](#), [3](#)
- [18] Michael Meyer, Georg Kuschik, and Sven Tomforde. Graph convolutional networks for 3d object detection on radar data. in 2021 iee. In *CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3053–3062, 2021. [7](#), [8](#)
- [19] Zhijun Pan, Fangqiang Ding, Hantao Zhong, and Chris Xiaoxuan Lu. Moving object detection and tracking with 4d radar point cloud. *arXiv preprint arXiv:2309.09737*, 2023. [1](#)
- [20] Loveneet Saini, Yu Su, Hasan Tercan, and Tobias Meisen. Centerpoint transformer for bev object detection with automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4451–4460, 2024. [3](#), [4](#), [7](#), [8](#)
- [21] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather, 2021. [2](#)
- [22] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. [3](#)
- [23] Peter Svenningsson, Francesco Fioranelli, and Alexander Yarovoy. Radar-pointgcn: Graph based object recognition for unstructured radar point-cloud data. In *2021 IEEE Radar Conference (RadarConf21)*, pages 1–6. IEEE, 2021. [1](#), [7](#), [8](#)
- [24] Bin Tan, Zhixiong Ma, Xichan Zhu, Sen Li, Lianqing Zheng, Sihan Chen, Libo Huang, and Jie Bai. 3d object detection for multi-frame 4d automotive millimeter-wave radar point cloud. *IEEE Sensors Journal*, 2022. [4](#)
- [25] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. [3](#), [4](#)
- [26] Michael Ulrich, Sascha Braun, Daniel Köhler, Daniel Niederlöhner, Florian Faion, Claudius Gläser, and Holger Blume. Improved orientation estimation and detection with hybrid object detection networks for automotive radar, 2022. [3](#), [4](#), [6](#), [7](#)

- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [1](#)
- [28] Baowei Xu, Xinyu Zhang, Li Wang, Xiaomei Hu, Zhiwei Li, Shuyue Pan, Jun Li, and Yongqiang Deng. Rpfa-net: A 4d radar pillar feature attention network for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3061–3066. IEEE, 2021. [4](#), [7](#)
- [29] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. [7](#)
- [30] Ryoma Yataka, Pu Wang, Petros Boufounos, and Ryuhei Takahashi. Radar perception with scalable connective temporal relations for autonomous driving. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13266–13270. IEEE, 2024. [2](#), [3](#), [7](#)
- [31] Ryoma Yataka, Pu Wang, Petros Boufounos, and Ryuhei Takahashi. Sira: Scalable inter-frame relation and association for radar perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15024–15034, 2024. [3](#), [7](#)
- [32] Ao Zhang, Farzan Erlik Nowruzzi, and Robert Laganieri. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102. IEEE, 2021. [1](#), [3](#)
- [33] Taohua Zhou, Mengmeng Yang, Kun Jiang, Henry Wong, and Diange Yang. Mmw radar-based technologies in autonomous driving: A review. *Sensors*, 20(24):7283, 2020. [1](#), [2](#)
- [34] Yi Zhou, Lulu Liu, Haocheng Zhao, Miguel López-Benítez, Limin Yu, and Yutao Yue. Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. *Sensors*, 22(11):4208, 2022. [1](#), [2](#)
- [35] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. [6](#)