MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

# RADAR PERCEPTION WITH SCALABLE CONNECTIVE TEMPORAL RELATIONS FOR AUTONOMOUS DRIVING

Yataka, Ryoma; Wang, Pu; Boufounos, Petros T.; Takahashi, Ryuhei

TR2024-023 March 19, 2024

**Abstract**

Due to the noise and low spatial resolution in automotive radar data, exploring temporal relations of learnable features over consecutive 2 radar frames has shown performance gain on downstream tasks (e.g., object detection and tracking) in our previous study [1]. In this paper, we further enhance radar perception by significantly extending the time horizon of temporal relations. To this end, we propose a scalable connective temporal radar (SCTR) method that consists of 1) a standard temporal relation layer (TRL), 2) a connective TRL with shifted window attention, and 3) a window merging operation, to facilitate feature connectivity between radar frames over an extended time interval. Our complexity analysis and comprehensive evaluation of the Radiate dataset demonstrate that the SCTR achieves a great tradeoff between the complexity and downstream detection performance.

*IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024*



Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

# RADAR PERCEPTION WITH SCALABLE CONNECTIVE TEMPORAL RELATIONS FOR AUTONOMOUS DRIVING

*Ryoma Yataka*[1,2], *Pu Wang*[1], *Petros Boufounos*[1], *and Ryuhei Takahashi*[2]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
[2]Information Technology R&D Center, Mitsubishi Electric Corporation, Kanagawa 247-8501, Japan

## ABSTRACT

Due to the noise and low spatial resolution in automotive radar data, exploring temporal relations of learnable features over consecutive 2 radar frames has shown performance gain on downstream tasks (e.g., object detection and tracking) in our previous study [1]. In this paper, we further enhance radar perception by significantly extending the time horizon of temporal relations. To this end, we propose a scalable connective temporal radar (SCTR) method that consists of 1) a standard temporal relation layer (TRL), 2) a connective TRL with shifted window attention, and 3) a window merging operation, to facilitate feature connectivity between radar frames over an extended time interval. Our complexity analysis and comprehensive evaluation of the Radiate dataset demonstrate that the SCTR achieves a great tradeoff between the complexity and downstream detection performance.

***Index Terms***— Autonomous driving, automotive perception, ADAS, radars, object detection, temporal attention.

## 1. INTRODUCTION

Automotive perception entails the interpretation of external driving surroundings and internal vehicle cabin conditions with an array of perception sensors to achieve robust safety and driving autonomy [2]. Compared with the mainstream camera and Lidar sensors, radar is cost-efficient, friendly to sensor maintenance and calibration, and has distinctive advantages in providing long-range perception capabilities in adverse weather and light conditions [3].

Nevertheless, a notable limitation of radar-assisted automotive perception is the low angular resolution in the azimuth and elevation domains and the inherent noise including multipath and ghost reflection. Consequently, its capacity for object detection and tracking lags behind the requirements for fully autonomous driving capabilities. Compared with radar-assisted multimodal automotive perception [4–6], standalone radar-only perception has been studied in [1, 7–12]. For instance, Bai et al. [10] introduced a radar transformer that uses vector and scalar attention mechanisms to establish attention maps across spatial, Doppler, and radar cross-section (RCS) domains using only radar information. This approach requires fewer perception resources and circumvents the need for intricate synchronization processes among multimodal sensors. A multi-view feature fusion method was proposed in [9] to combine features from range-Doppler, range-angle, and angle-Doppler radar heatmaps for object classification. As opposed to single-frame radar feature extraction, our previous study in [1] proposed a framework referred to as *TempoRadar* (TR) to explore temporal attention over features from 2 consecutive radar frames. It has shown promising performance gain evaluated using the large-scale open Radiate dataset.

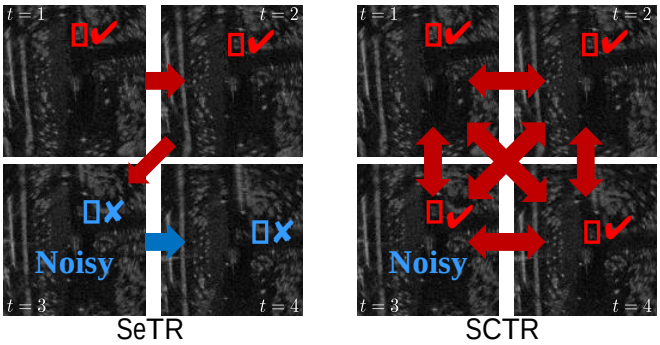


**Fig. 1**: Scalable Connective TempoRadar (SCTR). Compared to Sequential TR (SeTR), which is a sequential extension of vanilla TR, SCTR, which correlates with all time points in both directions, is robust to noise in radar frames.

One might then postulate: "What are the implications of extending TempoRadar to cover more radar frames?" The answer might be two-fold. On one hand, one should expect improved performance under the assumption that most radar features are present over more than just 2 frames, considering a typical radar frame rate of over 10 fps. On the other hand, directly applying temporal attention to a longer time horizon incurs a quadratic computation complexity over the number of features from each frame and the number of frames, compromising its scalability. One straightforward way for a scalable TempoRadar is to stack temporal feature attention for two consecutive frames and sequentially connect them, which we refer to as *sequential TR* (SeTR). However, the SeTR baseline may suffer from the noise, object occlusion, and slow convergence to capture temporal relation over a long time horizon as shown in the left plot of Fig.1.

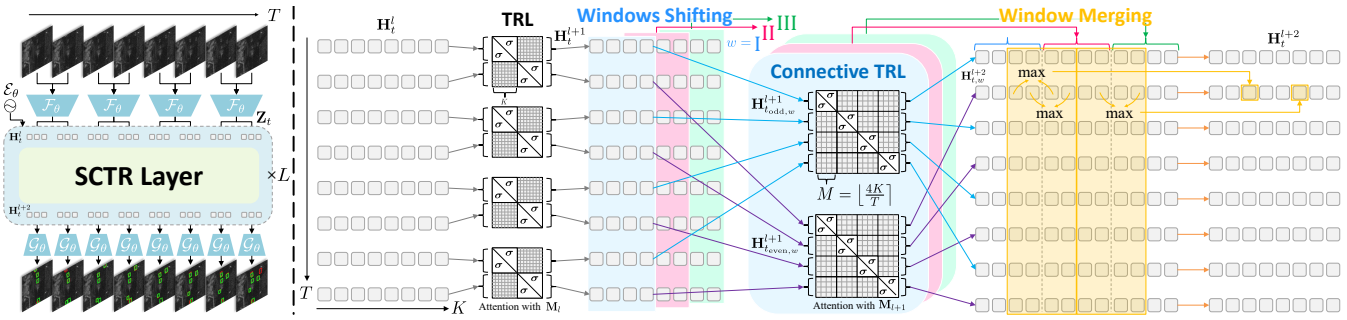


**Fig. 2**: (Left) Overall SCTR framework and (Right) SCTR Layer. This is constructed with three kinds of specific layers, Those are the vanilla temporal relation layer (TRL), the connective TRL and the window merging.

In this paper, we propose a *scalable connective Tempo-Radar* (SCTR) method that enables scalable inference over consecutive radar frames at a long time horizon (e.g., 8 frames) while facilitating connective attention across all considered frames; see the right plot of Fig.1. This is achieved by using the standard temporal relation layer (TRL) of [1] as a base layer, then partitioning top-$K$ features from the TRL outputs into shifted overlapping windows, designing attention mechanisms to maintain connectivity across time frames, and a window merging operation. Through our complexity analysis and performance evaluation on the Radiate dataset [13], the proposed SCTR shows competitive object detection performance with affordable complexity over a list of baseline methods including the TR and SeTR.

## 2. RADAR PERCEPTION WITH TEMPORAL RELATIONS FRAMEWORK

Denote multiple radar frames as $\mathbf{I} = \{I_1, \cdots, I_t, \cdots I_T\} \in \mathbb{R}^{T\times H\times W}$ where $I_t \in \mathbb{R}^{1\times H\times W}$ is a radar frame, and $T$, $H$ and $W$ represent the number of frames, height and width, respectively. Following the procedures of [1], we obtain the feature representations $\mathbf{Z}_t := \mathcal{F}_\theta(\mathbf{I}_{t,t-1})$ with a backbone $\mathcal{F}_\theta(\cdot)$. It is built in standard convolutional neural networks such as ResNet [14] which has shared model parameters $\theta$ for processing an input $\mathbf{I}_{t,t-1}$, which $\mathbf{I}_{t,t-1} = \{I_t, I_{t-1}\}$ if the $t$ is even and $\mathbf{I}_{t,t-1} = \{I_{t-1}, I_t\}$ if $t$ is odd. The final representations of the features result in $\mathbf{Z}_t \in \mathbb{R}^{C\times\frac{H}{s}\times\frac{W}{s}}$ where $C$ is the number of channels and $s$ is the downsampling ratio over the spatial dimension.

To localize objects, the 2D coordinates of the top-$K$ peak values in the heatmap, those width & length, orientation, and offset for compensating for the shifts are predicted as the bounding box of an object. The heatmap is obtained by a filtering module $\mathcal{G}^{\text{hm}}_\theta : \mathbb{R}^{C\times\frac{H}{s}\times\frac{W}{s}} \to \mathbb{R}^{1\times\frac{H}{s}\times\frac{W}{s}} : \mathbf{Z}_t \mapsto \mathbf{Z}^{\text{hm}}_t$ with parameters $\theta$ that is built with standard fully connected neural networks (FNN), and it is followed by a sigmoid function. The width $w$ & length $h$ are predicted through a head $\mathcal{G}^{\text{b}}_\theta : \mathbf{Z}_t\left[P^{\text{hm}}_t\right] \mapsto (w, h)$ for each top-$K$ features, where $P^{\text{hm}}_t$ is the set of coordinates defined as follows:

$$P^{\text{hm}}_t := \left\{(x,y) \;\middle|\; \mathcal{G}^{\text{hm}}_\theta(\mathbf{Z}_t)_{xy} \geq \left[\mathcal{G}^{\text{hm}}_\theta(\mathbf{Z}_t)\right]_K\right\}, \quad (1)$$

where $\left[\mathcal{G}^{\text{hm}}_\theta(\mathbf{Z}_t)\right]_K$ is the $K$-th largest value in $\mathcal{G}^{\text{hm}}_\theta(\mathbf{Z}_t)$ over the spatial space $\frac{H}{s}\times\frac{W}{s}$, and the subscript $xy$ denotes taking value at coordinate $(x, y)$. The orientation $\vartheta = \tan^{-1}\left(\frac{\sin\vartheta}{\cos\vartheta}\right)$ is predicted through the orientation head $\mathcal{G}^{\text{r}}_\theta : \mathbf{Z}_t\left[P^{\text{hm}}_t\right] \mapsto (\cos\vartheta, \sin\vartheta)$, and the offset is obtained by a offset head $\mathcal{G}^{\text{o}}_\theta : \mathbf{Z}_t\left[P^{\text{hm}}_t\right] \mapsto (o_x, o_y)$.

## 3. SCALABLE CONNECTIVE TEMPORALITY

We show the SCTR framework on the left side of Fig.2. SCTR employs an encoder to transform input into high-level features. The SCTR layer accentuates object positions. Then, the decoder estimates the bounding boxes. Subsequent subsections detail the constituents of the SCTR layer (Fig.2).

### 3.1. Temporal Relational Layer

TRL receives multiple feature vectors from the two frames. By taking the coordinate sets $P^{\text{pre-hm}}_t$, which is obtained from (1) with $\mathcal{G}^{\text{pre-hm}}_\theta \in \mathbb{R}^{C\times\frac{H}{s}\times\frac{W}{s}} \to \mathbb{R}^{1\times\frac{H}{s}\times\frac{W}{s}}$, into feature representations, we have the selective top-$K$ features matrix $\mathbf{H}_t := \mathbf{Z}_t\left[P^{\text{pre-hm}}_t\right] \in \mathbb{R}^{C\times K}$. Let $\mathbf{H}_{t,t-1} := \{\mathbf{H}_t, \mathbf{H}_{t-1}\}^\top \in \mathbb{R}^{2K\times C}$ denote the matrix concatenation that forms the input to the TRL. Furthermore, the positional encoding $\mathbf{P}^{\text{enc}}_t = \mathcal{E}_\theta\left(P^{\text{pre-hm}}_t\right) \in \mathbb{R}^{K\times D_{\text{pos}}}$ in the feature vectors, which $\mathcal{E}_\theta(\cdot)$ is built in FNN, is used before passing $\mathbf{H}_{t,t-1}$ into the TRL, so we get $\mathbf{H}^{\text{pos}}_{t,t-1} = \left\{\mathbf{H}_{t,t-1}, \mathbf{P}^{\text{enc}}_{t,t-1}\right\} \in \mathbb{R}^{2K\times(C+D_{\text{pos}})}$. In computing the TRL, we follow [15–18] by including a temporal inductive bias with a masking matrix $\mathbf{M}$ to each head:

$$\mathcal{A}(\mathbf{V}, \mathbf{X}) := \text{softmax}\left(\frac{\mathbf{M} + q(\mathbf{X})\,k(\mathbf{X})^\top}{\sqrt{d}}\right) v(\mathbf{V}), \quad (2)$$

where $q(\cdot)$, $k(\cdot)$ and $v(\cdot)$ are linear transformation layers and are referred as query, keys and values respectively. $d$ is the

dimension of the query and the keys and is used to scale the dot product between them. In the TRL of layer $l$, we obtain the attention $\mathcal{A}\left(\mathbf{H}_{t,t-1}^{l}, \mathbf{H}_{t,t-1}^{l,\text{pos}}\right)$ with the following masking matrix $\mathbf{M}_l$:

$$\mathbf{M}_l := \begin{bmatrix} \mathbb{I}_K, \mathbf{1}_K \\ \mathbf{1}_K, \mathbb{I}_K \end{bmatrix} + \sigma \left( \begin{bmatrix} \mathbf{1}_K, \mathbf{0}_K \\ \mathbf{0}_K, \mathbf{1}_K \end{bmatrix} - \mathbb{I}_{2K} \right), \quad (3)$$

where a block $\mathbb{I}_K$ is the identity matrix of size $K$, $\mathbf{1}_K$ and $\mathbf{0}_K$ are the all-one and all-zero matrix with size $K \times K$ respectively, and $\sigma$ is a negative constant that is set to $-10^{10}$ in our implementation to guarantee a near-zero value in the output through softmax. Diagonal blocks disable attention between features of the same frame, while off-diagonal blocks allow for cross-frame attention.

### 3.2. Connective TRL

CTRL uses the window shifting technique employed in the Swin Transformer [19] to establish interframe connections for each time step, while mitigating computational complexity. This allows the CTRL to achieve high inference performance even when radar frames containing low SNR are present. Initially, we divide the top-$K$ selected feature vectors calculated at the $l$-th layer into $\Omega = \left\lfloor \left\lceil \frac{K}{S} \right\rceil - \frac{M}{S} \right\rceil + 1$ windows using a window shifting with a stride $S = \left\lfloor \frac{M}{2} \right\rceil$, where $\lfloor \cdot \rceil$ denotes the round function and $M = \left\lfloor \frac{4K}{T} \right\rceil$ denotes the size of the shifted window. Subsequently, for each window $w = \mathrm{I}, \mathrm{II}, \cdots, \Omega$, the odd indexed frames and even indexed frames in the temporal direction are separated to create two subsets $\mathbf{H}_{t_{\text{even},w}}^{l+1}, \mathbf{H}_{t_{\text{odd},w}}^{l+1} \in \mathbb{R}^{\frac{TM}{2} \times C}$. The above division process ensures that the computational complexity of the attention remains at a low level and is scalable for the time domain, since the size of the attention is $\frac{TM}{2} \times \frac{TM}{2} \approx 2K \times 2K$. In CTRL, we apply (2) to each of these subsets to obtain attentions $\mathcal{A}\left(\mathbf{H}_{t_{\text{even/odd},w}}^{l+1}, \mathbf{H}_{t_{\text{even/odd},w}}^{l+1,\text{pos}}\right)$, where $\mathbf{H}_{t_{\text{even/odd},w}}^{l+1,\text{pos}} = \left\{\mathbf{H}_{t_{\text{even/odd},w}}^{l+1}, \mathbf{P}_{t_{\text{even/odd},w}}^{\text{enc}}\right\}$, with $\mathbf{M}_{l+1}$:

$$\mathbf{M}_{l+1} := \mathbf{B} + \sigma\left(\overline{\mathbf{B}}\right) \text{ s.t. } \mathbf{B} = \begin{bmatrix} \mathbb{I}_M & & \mathbf{1} \\ & \ddots & \\ \mathbf{1} & & \mathbb{I}_M \end{bmatrix}, \quad (4)$$

where $\overline{\mathbf{B}}$ is formed by changing the values of 1 within set $\mathbf{B}$ to 0, while simultaneously inverting the values of 1 to 0. Next, we sequentially apply a feed-forward function that consists of two linear layers, layer normalization and shortcut on features. The relational modeling is built with multiple SCTR layers with identical design. We obtain the updated features $\mathbf{H}_{t,w}^{l+2} \in \mathbb{R}^{C \times M}$ by dividing $\mathbf{H}_{t_{\text{even}},w}^{l+2\top}$ and $\mathbf{H}_{t_{\text{odd}},w}^{l+2\top}$.

### 3.3. Window Merging

Due to the use of window shifting, there exists overlap in the updated feature vectors. Consequently, feature vectors in overlapped positions must be integrated. Various integration methods such as summation or maximization can be considered, and in this paper, based on preliminary experiments, we have opted to use maximization. Applying the integration, we obtain the features $\mathbf{H}_t^{l+2} \in \mathbb{R}^{C \times K}$. Finally, we refill the feature vector to $\mathbf{Z}_t$ in the corresponding spatial coordinates of $P_t^{\text{pre-hm}}$.

### 3.4. Learning

We pick the object's center coordinates from the heatmap, and learn its attributes from feature representations through regression. Regression functions, which are heatmap loss $\mathcal{L}_t^{\text{h}}$, width & Length loss $\mathcal{L}_t^{\text{b}}$, orientation loss $\mathcal{L}_t^{\text{r}}$, and offset loss $\mathcal{L}_t^{\text{o}}$, compose the training objective by a linear combination:

$$\min_{\theta} \mathcal{L} := \sum_{t=1}^{T} \left\{ \frac{1}{N_{\text{gt}}} \sum_{k=1}^{N_{\text{gt}}} \left( \mathcal{L}_{t,k}^{\text{b}} + \mathcal{L}_{t,k}^{\text{r}} + \mathcal{L}_{t,k}^{\text{o}} \right) - \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{t,i}^{\text{h}} \right\}, \quad (5)$$

where $k$ denotes $k$-th ground truth object and the total number is $N_{\text{gt}}$, and $N$ denotes the total number of pixels in the heatmap. Each loss is as follows:

$$\begin{aligned} \mathcal{L}_{t,i}^{\text{h}} :=& \mathbb{1}_{\{h_{t,i}=1\}} \left(1 - \hat{h}_{t,i}\right)^{\alpha} \log\left(\hat{h}_{t,i}\right) \\ &+ \mathbb{1}_{\{h_{t,i}\neq 1\}} \left(1 - h_{t,i}\right)^{\beta} \hat{h}_{t,i}^{\alpha} \log\left(1 - \hat{h}_{t,i}\right), \quad (6) \end{aligned}$$

$$\mathcal{L}_{t,k}^{\text{b}} := S_{L_1}\left( \left\| \mathcal{G}_{\theta}^{\text{b}}\left(\mathbf{Z}_t\left[P_{t,k}^{\text{gt}}\right]\right) - \left(w_{t,k}, h_{t,k}\right)^{\top} \right\| \right), \quad (7)$$

$$\mathcal{L}_{t,k}^{\text{r}} := S_{L_1}\left( \left\| \mathcal{G}_{\theta}^{\text{r}}\left(\mathbf{Z}_t\left[P_{t,k}^{\text{gt}}\right]\right) - \left(\cos\vartheta_{t,k}, \sin\vartheta_{t,k}\right)^{\top} \right\| \right), \quad (8)$$

$$\mathcal{L}_{t,k}^{\text{o}} := S_{L_1}\left( \left\| \mathcal{G}_{\theta}^{\text{o}}\left(\mathbf{Z}_t\left[P_{t,k}^{\text{gt}}\right]\right) - \left(o_{x,t,k}, o_{y,t,k}\right)^{\top} \right\| \right), \quad (9)$$

where $h_{t,i}$ and $\hat{h}_{t,i}$ denote the ground-truth and predicted value at $i$-th coordinate in $\mathbf{Z}_t^{\text{hm}}$, and $\alpha$ and $\beta$ are hyperparameters and are chosen empirically with 2 and 4, respectively. $P_{t,k}^{\text{gt}}$ denotes the coordinate $(c_{x,t,k}, c_{y,t,k})$ of the center of $k$-th ground truth object, $(w_{t,k}, h_{t,k})$ is the width & length, $(o_{x,t,k}, o_{y,t,k}) = \left(c_{x,t,k}/s - \lfloor c_{x,t,k}/s \rceil, c_{y,t,k}/s - \lfloor c_{y,t,k}/s \rceil\right)$, and $S_{L_1}(\cdot)$ is a smooth $L_1$ loss [20]. For each training step, our training procedure calculates $\mathcal{L}$ and does the backward for both $t = 1$ to $t = T$ and $t = T$ to $t = 1$ simultaneously. Therefore, optimization can be viewed as a bidirectional backward-forward training through $T$ frames.

### 3.5. Complexity Analysis

The computational complexity of SCTR depends on the number of frames $T$ and $K$, and the number of layers $L$. Compared to vanilla TR, SeTR and SCTR have less complexities as follows:

$$\text{TR}: \quad (TK)^2 L \cdots \mathcal{O}\left(T^2K^2\right), \quad (10)$$

$$\text{SeTR}: \quad 4K^2 (T-1) L \cdots \mathcal{O}\left(K^2\right), \quad (11)$$

$$\text{SCTR}: 2K^2 (3T-4) L \cdots \mathcal{O}\left(K^2\right). \quad (12)$$

**Table 1**: Experimental results of object detection on *Radiate* dataset.

| | Split: train good weather | | | Split: train good and bad weather | | |
|---|---|---|---|---|---|---|
| | mAP@0.3 | mAP@0.5 | mAP@0.7 | mAP@0.3 | mAP@0.5 | mAP@0.7 |
| BBAVectors-ResNet18 | 59.38± 3.47 | 50.53± 2.07 | 19.72± 1.10 | 56.84± 3.45 | 45.43± 2.87 | 15.07± 1.76 |
| BBAVectors-ResNet34 | 60.88± 1.79 | 51.26± 1.99 | 19.86± 1.36 | 55.87± 2.90 | 44.61± 2.57 | 14.67± 1.45 |
| TR-ResNet18 2 frames | 62.79± 2.01 | 53.11± 1.96 | 20.57± 1.47 | 58.87± 3.31 | 46.42± 3.24 | 15.59± 2.31 |
| TR-ResNet34 2 frames | 63.63± 2.08 | 54.00± 2.16 | 21.08± 1.66 | 56.18± 4.27 | 43.98± 3.75 | 14.35± 2.15 |
| TR-ResNet18 4 frames | 66.37± 1.62 | 53.23± 1.67 | 19.59± 0.78 | 65.10± 1.67 | 52.47± 1.21 | 19.62± 1.33 |
| TR-ResNet34 4 frames | 67.48± 0.94 | 57.01± 1.03 | 22.46± 1.76 | 64.60± 2.08 | 51.99± 1.94 | 19.03± 1.10 |
| SeTR-ResNet18 4 frames | 65.97± 2.03 | 55.79± 2.12 | 21.90± 1.12 | 64.62± 1.79 | 51.78± 1.81 | **19.65**± 0.84 |
| SeTR-ResNet34 4 frames | 67.30± 1.80 | 56.61± 1.83 | 21.68± 1.24 | 65.51± 1.52 | 52.43± 1.51 | 19.63± 1.29 |
| SCTR-ResNet18 4 frames | 66.08± 2.52 | 55.45± 2.07 | 21.49± 1.19 | 65.16± 2.50 | 52.55± 2.14 | 19.22± 0.95 |
| SCTR-ResNet34 4 frames | **68.06**± 1.60 | **57.03**± 1.34 | **22.62**± 1.18 | **66.01**± 1.05 | **52.55**± 0.96 | 19.18± 1.02 |

**Table 2**: Ablation study of various number of frames $T$.

| Split: train good weather | mAP@0.3 | mAP@0.5 | mAP@0.7 |
|---|---|---|---|
| TR-ResNet34 6 frames | 65.33± 1.72 | 54.29± 1.56 | 20.56± 1.14 |
| SeTR-ResNet34 6 frames | 66.81± 2.04 | 56.16± 2.35 | **22.37**± 1.39 |
| SCTR-ResNet34 6 frames | **68.22**± 1.23 | **56.53**± 1.06 | 21.01± 1.12 |
| TR-ResNet34 8 frames | 65.89± 3.06 | 55.16± 3.04 | 21.04± 1.63 |
| SeTR-ResNet34 8 frames | 66.33± 2.39 | 55.98± 2.11 | **22.05**± 1.60 |
| SCTR-ResNet34 8 frames | **67.67**± 1.18 | **56.47**± 1.54 | 20.68± 2.19 |
| TR-ResNet34 10 frames | 65.78± 2.08 | 55.05± 1.72 | **21.25**± 1.42 |
| SeTR-ResNet34 10 frames | 66.78± 2.22 | 55.43± 2.11 | 20.91± 2.08 |
| SCTR-ResNet34 10 frames | **67.17**± 1.57 | **56.07**± 1.61 | 21.24± 0.93 |

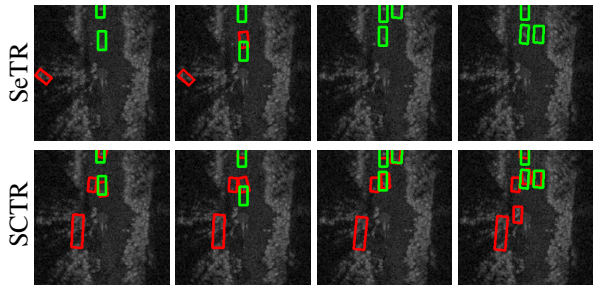


**Fig. 3**: Visualizations on Rain-4-0 in Radiate, with green denoting the ground truth and red denoting the predictions.

## 4. EXPERIMENTS

We use the radar dataset: Radiate [13] in our experiments as same as Li et al. [1]. It consists of video sequences recorded in adverse weathers, including sun, night, rain, fog and snow, and adopts the mechanical scan of the Navtech CTS350-X, which employs a FMCW radar, providing 360° range-azimuth images at 4 Hz. We follow the official 3 splits: train in good weather (22383 frames), train good and bad weather (9749 frames), and test (11305 frames, all kinds of weather conditions).

We implemented several detectors that have been well demonstrated in object detection for comparison. These detectors include the following: BBAVectors [21], vanilla TR [1] and SeTR that stacks self-attention for two frames and sequentially connects them through $T$ frames. Comparison is carried out with different numbers of layers with the ResNet backbone. For all numerical results, we apply a center crop with size $256 \times 256$ upon input images and exclude the targets outside this scope. The position dimension $D_{\text{pos}} = 64$, $s = 4$, $K = 8$, $L = 2$, the batch size is 16, the learning rate is 5e-4, and weight decay is 1e-2 for Adam optimizer with ten training epochs. We adopt mean Average Precision (mAP) with thresholds.

### 4.1. Result

We report the detection results in Tables 1 and 2. First, by using more than four frames, each method consistently outperforms the conventional approach that considers only two frames in both training splits with different IoU thresholds. In particular, our SCTR achieves outstanding mAP among them. Moreover, while the performance of TR and SeTR deteriorates with an increase in the number of frames $T$ beyond six, SCTR demonstrates a performance improvement even for longer frame sequences. In contrast, SeTR outperforms the other two methods in terms of mAP@0.7. This is attributed to its ability to fine-tune the object's position by sequentially tracking vehicle features. Fig.3 shows the visualization results from bad weather: Rain-4-0, with green denoting the ground truth and red denoting the predictions. Our SCTR outperforms SeTR in accurate vehicle detection. However, alongside correct predictions, our model introduces false positives, often appearing as clusters of reflections resembling ghost objects within boxes. Reducing false positives is an intriguing challenge for future exploration.

## 5. CONCLUSION

In the realm of autonomous driving, we focused on object detection using radar. Our approach aimed to improve radar perception by integrating temporal information from multiple frames. To achieve this, we introduced a comprehensive SCTR framework that captures and models object-level coherence. This layer effectively captured and modeled object-level coherence across extended time frames. The experimental results in object detection provided evidence of the effectiveness of our approach.

## 6. REFERENCES


[1] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17050–17059.

[2] A. Pandharipande, C. H. Cheng, J. Dauwels, S. Z. Gurbuz, J. I. Guzman, G. Li, A. Piazzoni, P. Wang, and A. Santra, "Sensing and machine learning for automotive perception: A review," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 11097–11115, 2023.

[3] S. Zeng and J. N. Nickolaou, *Automotive Radar*, CRC Press, 03 2014.

[4] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 444–453.

[5] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: Exploiting radar for robust perception of dynamic objects," in *Computer Vision – ECCV 2020*, 2020, pp. 496–512.

[6] T. Y. Lim, A. Ansari, B. Major, D. Fontijne, M. Hamilton, R. Gowaikar, and S. Subramanian, "Radar and camera early fusion for vehicle detection in advanced driver assistance systems," in *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on NeurIPS*, 2019.

[7] A. Zhang, F. E. Nowruzi, and R. Laganiere, "RADDet: Range-azimuth-doppler based radar object detection for dynamic road users," in *18th Conference on Robots and Vision*, 2021, pp. 95–102.

[8] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut, "Multi-view radar semantic segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15651–15660.

[9] X. Gao, G. Xing, S. Roy, and H. Liu, "RAMP-CNN: A novel neural network for enhanced automotive radar object recognition," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5119–5132, 2021.

[10] J. Bai, L. Zheng, S. Li, B. Tan, S. Chen, and L. Huang, "Radar Transformer: An object classification network based on 4d mmw imaging radar," *Sensors*, vol. 21, no. 11, 2021.

[11] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[12] R. Zheng, S. Sun, H. Liu, and T. Wu, "Deep neural network enabled vehicle detection using high-resolution automotive radar imaging," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 5, pp. 4815–4830, 2023.

[13] M. Sheeny, E. D. Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A radar dataset for automotive perception in bad weather," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 1–7.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, 2020.

[16] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, S. Piao, J. Gao, M. Zhou, and H. Hon, "UNILMv2: Pseudo-masked language models for unified language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[17] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588–3597.

[18] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3463–3472.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9992–10002.

[20] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[21] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 2149–2158.