TFoC: A Feature Gimbal for Radar Semantic Segmentation

Anonymous authors
Paper under double-blind review

Abstract

The proper utilization of spatio-temporal information is essential for radar semantic segmentation (RSS). Classical 3D convolutions or spatio-temporal attention mechanisms can help an RSS model preserve spatio-temporal information to a certain extent. However, since the objects of interest in radar applications are dynamically moving, these "*naïve*" computational modules are not the optimal choice in the absence of a spatio-temporal alignment mechanism. To address this, we propose TFoC (Temporal Focusing Convolution), motivated by the working principle of a camera gimbal, *i.e.*, dynamically adjusting the lens pose to steadily focus on the region of interest. Inspired by this, the receptive field design of TFoC aims to make spatio-temporal convolution operations focused rather than blind, and the sampling efficiency in the spatio-temporal domain dense rather than sparse. Therefore, TFoC is fundamentally a novel learnable module inspired by physical intuition. To better elucidate the learning mechanism of TFoC, we start from the classical Maximum A Posteriori (MAP) estimation and incorporate Bayesian theory to provide an interpretive perspective for the design of TFoC.

1 Introduction

Radar, based on the principle of electromagnetic wave reflection Richards (2005), serves as an important device for detecting the motion state of objects. It owns better robustness against bad weather Bijelic et al. (2018) than cameras, and higher resolution in both location and angle fields than Lidar. Meanwhile, radar echo can be conveniently converted into two-dimensional frequency representations (*e.g.*, Range-Doppler, Range-Angle) via the Fourier transform, enabling adaptive, data-driven feature extraction by deep networks Brodeski et al. (2019); Major et al. (2019); Wang et al. (2021); Gao et al. (2021), free from the strong prior assumptions and heavy manual tuning of classical detectors. Up to now, learnable radar perception has gradually become a hot research topic in the field of radar detection.

Among the research directions of radar perception, Radar Semantic Segmentation (RSS) has been studied most sufficiently and deeply. Early explorations integrated the modules and architectures repeatedly verified in computer vision into the design of RSS models, achieving obviously better detection performance than traditional detection methods. RSS-Net Kaul et al. (2020) is the first end-to-end RSS network. MV-Net Ouaknine et al. (2021a) further proposed a multi-view RSS network and combined it with the multi-scale pyramid strategy to obtain unit-level semantic segmentation in RD and RA views. Compared with target detection methods with detection anchors Ren et al. (2017); Redmon et al. (2016), such RSS models can achieve fine-grained detection while identifying targets and background simultaneously. However, radar signals **do not own such rich semantic information as images**, *e.g.*, color, structure, and texture. Methods proven effective in vision therefore encounter a performance ceiling in radar perception Long et al. (2015); Ronneberger et al. (2015); Chen et al. (2017); Sun et al. (2019), and further improvement calls for designs tailored to the characteristics of radar data. Typical examples include PKC Zhang et al. (2023b) and AdaPKC Li et al. (2024), which focus on mining the saliency of targets within a single CPI, *i.e.*, the spatial dimension, by introducing the classic radar principle CFAR Rohling (1983) into deep learning operators.

In addition to target saliency, TARSS-Net Zhang et al. (2024) has also demonstrated that effectively utilizing the **spatio-temporal information of radar data** is another effective way to improve RSS performance. Nevertheless, as a detection device for long-range moving targets, the target information

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in radar data is often extremely sparse Zhang et al. (2023a), especially for typical UAV detection data such as KuRALS Li et al. (2026), and the motion of the targets themselves further aggravates this sparsity in the spatio-temporal domain. So far, such extreme spatio-temporal sparsity of radar targets has not been fully studied, which we believe is the key to further improving the representation effectiveness of radar data.

We argue that the key to solving the spatio-temporal sparsity of targets lies in improving the semantic density within a single computation, or, finding a computation paradigm with semantic purification capability. Inspired by the working principle of a camera gimbal, *i.e.*, dynamically adjusting the lens pose to steadily focus on the region of interest, we propose a brand-new spatio-temporal representation module, **T**emporal **Fo**cusing **C**onvolution (TFoC), to track and aggregate the movement trajectory of the target in time for spatio-temporal feature focusing. Our contributions are as follows.

- ⋆ **A novel practical spatio-temporal focusing convolution paradigm**. To fully exploit the correlation of radar targets in spatio-temporal dimension, we first propose a radar-specific spatio-temporal focusing convolution, TFoC. It significantly improves the learning effectiveness by increasing semantic correlation within a single calculation, while keeping the computation simple and efficient. It can be freely applied to any network instead of traditional 2D/3D convolution.
- ⋆ **Temporal focusing method based on spatio-temporal trajectory probabilistic modeling**. In TFoC convolution diagram, consistency measurement and feature aggregation are two crucial points for constructing receptive field and aggregate correlation features, respectively. We innovatively achieve temporal focusing based on probability modeling of radar target motion trajectory in both spatial and temporal dimensions, which enhances target saliency and mines key information from cluttered and redundant radar signals.
- ⋆ **Supervising functions for spatio-temporal consistency**. In order to make full use of target motion information in historical data, Track Loss is proposed to strengthen the probability model constraint during the learning of spatial-temporal target trajectory.

2 Related works

Spatiotemporal information encompasses not only the spatial structures (e.g., object shapes, textures) of single-frame data but also the deep semantics embedded in inter-frame, such as temporal dependencies and dynamic interactions. We categorize existing spatiotemporal information learning methods into implicit spatiotemporal modeling and explicit spatiotemporal modeling, and clarify why they can hardly be the best answer to the extreme spatio-temporal sparsity of radar targets.

Implicit spatiotemporal modeling. These methods do not directly learn spatiotemporal information, but achieve its implicit utilization by adjusting feature flows within network layers. According to different temporal data processing approaches, this category can be further divided into two subgroups. i) *Causal Temporal Relation Modeling.* Prototyped by Hidden Markov Models (HMMs), this subgroup includes Recurrent Neural Networks (RNNs) Elman (1990), Long Short-Term Memory networks (LSTMs) Hochreiter & Schmidhuber (1997), and Gated Recurrent Units (GRUs) Cho et al. (2014), which propagate hidden states across time steps through recurrent connections. However, hidden states are updated by indiscriminately reading every frame, under which the scarce target evidence is easily diluted by massive clutter during long-term propagation. ii) *Parallelized Sequence Representation Modeling.* 3D convolution simultaneously aggregates local spatial and temporal information with 3D kernels, and is thus one of the few computer vision methods directly transferable to radar data Tran et al. (2015); Carreira & Zisserman (2017); Diba et al. (2017). Nevertheless, since radar targets occupy only a few units, most of the kernel positions actually aggregate the background, resulting in extremely low efficiency of semantic learning Chen et al. (2017); Sun et al. (2019). Transformers Bertasius et al. (2021); Arnab et al. (2021) segment grid data into tokens to achieve global spatiotemporal correlations, but at the cost of $O(T^2)$ complexity, which is prohibitive for large-scale long-span data (e.g. radar data) even with lightweight variants. Recent radar-oriented Transformers, including TransRSS Zou et al. (2023), TransRadar Dalbah et al. (2024), and T-RODNet Jiang et al. (2023), further demonstrate the benefit of self-attention for radar perception, but spatiotemporal attention measures pairwise feature affinity rather than the physical correspondence of the same moving target; under extreme sparsity, both the computation and the aggregation weights are dominated by massive target-irrelevant pairs, so the target evidence can hardly be concentrated.

Explicit spatiotemporal modeling. In contrast, explicit spatiotemporal modeling breaks the structured learning mechanism and directly models the correspondence between individual learning units across frames, which is suitable for unstructured data such as object motion modeling Chu et al. (2017) and traffic flow prediction Yao et al. (2018). Graph Neural Networks (GNN) Yan et al. (2018) and optical flow estimation Dosovitskiy et al. (2015); Teed & Deng (2020) are two typical methods of explicit spatiotemporal modeling. However, the lack of semantics in radar signals makes graph construction for GNNs difficult; sparse targets and severe interference, in turn, render optical flow estimation challenging. Neither of the two methods is applicable to radar signals.

In summary, existing methods aggregate temporal information either on **dense-and-fixed grids**, or through global but **content-agnostic pairing**, or in an implicit feature-flow manner. None of them explicitly concentrates the computation on the moving target so as to improve the semantic density of a single computation. Considering the characteristics of radar data, we propose a novel spatiotemporal information mining method, TFoC. Different from adaptive receptive-field operators such as dilated Yu & Koltun (2016) and deformable convolutions Dai et al. (2017); Zhu et al. (2019), which only adjust the spatial sampling locations of static kernels, TFoC is like **a feature gimbal** that can stably track target features in the spatio-temporal dimension and could further aggregate these useful features scattered across the timeline.

3 METHODOLOGY

Preliminaries. TFoC is inspired by physical intuition, while its concrete implementation is an engineering trade-off tailored to the RSS task. To elucidate the learning mechanism behind TFoC without obscuring its practical design, we describe the concrete implementation of TFoC in this section, and defer the complete theoretical interpretation, *i.e.*, the probabilistic derivations based on Maximum A Posteriori (MAP) estimation and Bayesian theory, to Appendix A. To distinguish between the **engineering implementation** and the **theoretical interpretation**, we adopt distinct notation systems for key variables: calligraphic and Greek letters are used more extensively in the theoretical sections, while standard alphabetic symbols are predominantly used in the implementation sections. Table 4 gives the basic notations for TFoC description.

TFoC is designed as a plug-and-play convolution-style learning module, which can be used just like any standard layer. However, unlike current spatio-temporal learning modules, TFoC is designed to span the entire RSS optimization: its input is provided by a local spatial compression (LSC) module, its output directly serves a dedicated loss function termed **Track Loss**, and the critical intermediate components handle **spatio-temporal semantic alignment-based focusing (SAF & TCA)**. The merit of this design is that it remains indistinguishable from standard learning modules in engineering practice, yet offers sufficient interpretability in terms of its underlying mechanism.

3.1 SEMANTIC ALIGNMENT-BASED SPATIAL FOCUSING (SAF)

Consider: if we could design a 3D convolution operation whose receptive field consistently remains focused on the objects of interest, such a receptive field would exhibit high semantic purity. Would this not significantly reduce the classification difficulty for the segmentation head? Semantic alignment through **SAF** follows precisely this intuition: it uses the features at the current timestep as an anchor to collect features from historical timesteps that share consistent semantic information, thereby constructing a spatio-temporal receptive field with high semantic purity.

3.1.1 MAP PERSPECTIVE ON SAF: A SUMMARY

The focusing behavior of SAF can be interpreted from a Maximum A Posteriori (MAP) perspective; we summarize the key conclusions here and defer the complete derivation to Appendix A.1. Given a target feature point $\mathbf{x}^\star$ and a historical candidate $\mathbf{x}^\dagger_{t,n}$, we define the **Semantic Alignment Probability (SAP)** $P(j^\star = j^\dagger_{t,n} \mid \mathbf{x}^\star, \mathbf{x}^\dagger_{t,n})$, *i.e.*, the posterior that the two points share the same semantic label ($j^\star$/$j^\dagger$ denotes semantic label for target/historical feature point). Under a conditional independence assumption, the SAP factorizes over classes into two posterior terms, both of which can be estimated directly by the RSS backbone. Maximizing the SAP within each Historic Search Field (HSF)

$\mathcal{H}_t \doteq \{\mathbf{x}_{t,n}^{\dagger}\}_{n=1}^{N}$ then yields a **hard selection** rule (Eq. 16) that retains, from every historical frame, exactly the candidate most semantically coherent with $\mathbf{x}^{\star}$, denoted as $\mathbf{x}_{\hat{n}_t}^{\dagger}$.

3.1.2 TARGET-HISTORY RELATION-WEIGHTED SOFT FOCUSING

While hard focusing via MAP offers more intuitive interpretability, the strict $\arg\max$ operation is non-differentiable and would block gradient flow through the entire RSS model, impairing end-to-end optimization. Hence, a simple weighted summation mechanism is adopted to interpolate a new historical candidate, softening the focusing process. Here, we draw inspiration from the *Target-History Relation* (THR) idea in Zhang et al. (2024), to redesign the local alignment operation.

THR as an Approximate SAP Estimator. From the MAP perspective, SAP estimation explicitly requires `SoftMax` logit prediction for the target and its historical candidate from RSS backbone, *i.e.*, $\mathbf{q}^{\star}, \mathbf{q}_{t,n}^{\dagger} \in \Delta^{J}$, forcing a restructuring of the RSS pipeline from single-input prediction to multi-input multi-prediction. Given the high dimensionality of radar spatio-temporal tensor data, this would introduce prohibitive memory overhead and computational burden. Even setting aside resource consumption, such a design poses significant engineering challenges for edge deployment. To avoid this, we instead learn the semantic relation weights, avoiding the explicit computation of $\mathbf{q}^{\star}, \mathbf{q}_{t,n}^{\dagger}$:

$$\text{THR as SAP: } A_{t,n} = \psi\left(\mathbf{x}^{\star}, \mathbf{x}_{t,n}^{\dagger};\ \boldsymbol{\omega}_{\text{SAF}}\right), \tag{1}$$

where $\psi(\cdot,\cdot;\ \boldsymbol{\omega}_{\text{SAF}}) : \mathbb{R}^{2C} \to \mathbb{R}$ can be viewed as a kernel function operating in the feature space obtained by the LSC, computing the semantic relation weight for each candidate feature pair of target-history. In the concrete implementation, each pairwise evaluation of $\psi(\cdot,\cdot;\ \boldsymbol{\omega}_{\text{SAF}})$ is performed by a single convolution operation.

Target-History Relation-Weighted Soft Focusing. In the MAP perspective, $\arg\max$ is used to locate $\mathbf{x}_{\hat{n}_t}^{\dagger}$, followed by an indexing operation to gather $\hat{\mathcal{H}}$. This forces a continuous, differentiable multiply-accumulate operation to be truncated by a hard selection (indexing) operation. To avoid this, a "softened" selection is needed to maintain continuous gradient flow. The simplest realization of such soft selection is a weighted summation with weights approximating a one-hot distribution. We, accordingly, propose the following design:

$$\boxed{\text{SAF: } \hat{\mathbf{H}} \doteq \left\{\hat{\mathbf{x}}_t^{\dagger}\right\}_{t=1}^{\tau} = \left\{\sum_{n=1}^{N} a_{t,n} \cdot \mathbf{x}_{t,n}^{\dagger}\right\}_{t=1}^{\tau}, \quad \text{where } a_{t,n} = \frac{\exp(A_{t,n}/\sigma_{\text{spa}})}{\sum_{m=1}^{N}\exp(A_{t,m}/\sigma_{\text{spa}})}.} \tag{2}$$

Here, σ_{spa} is a control factor that modulates the "hardness" of spatial focusing. As $\sigma_{\text{spa}} \to 1$, the distribution of $a_{t,n}$ becomes softer, equivalent to applying `SoftMax` directly over $A_{t,n}$; as $\sigma_{\text{spa}} \to 0^{+}$, the distribution of $a_{t,n}$ approaches a one-hot encoding, producing hard focusing behavior.

Consequently, we can adopt a "soft-to-hard" annealing strategy to gradually drive $\hat{\mathbf{H}}$ toward $\hat{\mathcal{H}}$, so that TFoC conforms to the hard selection mechanism under the MAP perspective as training converges.

3.2 TRAJECTORY CONSISTENCY-BASED TEMPORAL AGGREGATION (TCA)

In SAF, we achieved pairwise local spatial focusing through the target-history semantic relation metric. However, this suffers from two limitations: 1) The specific semantic meaning is folded (via marginalization over classes). 2) There is no consistency constraint across the full temporal sequence. These two limitations restrict the semantic purification capability of TFoC. This section addresses this remaining problem, which we summarize as **TCA**. Similarly, we adopt a dual perspective combining theoretical interpretation and engineering implementation to describe the TCA methodology.

3.2.1 BAYESIAN PERSPECTIVE ON TCA: A SUMMARY

TCA also admits a probabilistic interpretation, which we summarize here with full derivation deferred to Appendix A.2. We model trajectory consistency as a binary event $\mathcal{T}^{\star}$ indicating whether the focused historical sequence $\hat{\mathcal{H}}$ and the target point $\mathbf{x}^{\star}$ belong to the same motion trajectory, and take the Trajectory Consistency Posterior $P(\mathcal{T}^{\star} = 1|\hat{\mathcal{H}}, \mathbf{x}^{\star})$ (TCP in Eq. 19) as aggregation objective. Under

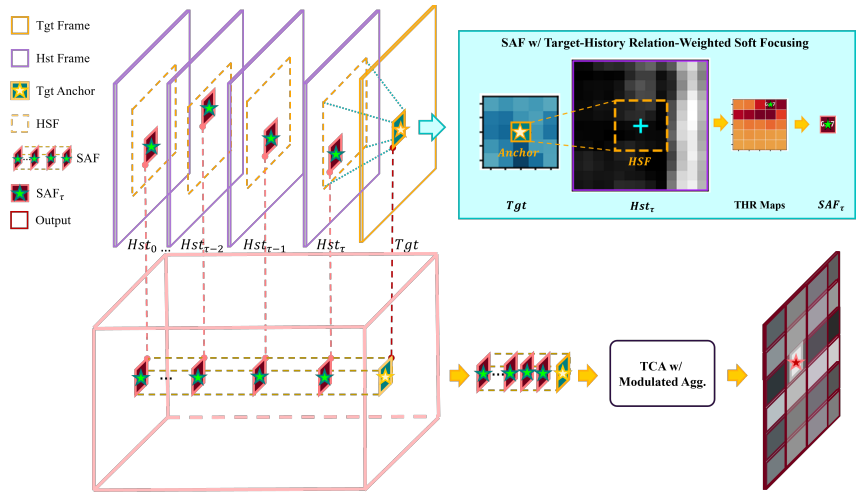


Figure 1: Schematic diagram of TFoC working principle: SAF for spatial alignment (Eq. 2) and TCA for temporal aggregation (Eq. 5).

a Markov assumption, trajectory-conditioned likelihood decomposes over historical frames. Since semantic alignment is a necessary (though not sufficient) condition for trajectory consistency, each per-frame factor can be approximated by the corresponding SAP, which serves as a **discriminative surrogate** for the trajectory likelihood van den Oord et al. (2018). Maximizing TCP is thereby equivalent to maximizing the joint of per-frame alignment probabilities $P(j^\star = j_t^\dagger | \mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger)$ and a class-dependent trajectory prior $P(\mathcal{T}^\star | \mathbf{x}^\star)$ (see Eq. 23).

3.2.2 Discriminative TCA with Track Loss

Similarly to SAF, strict MAP optimization involves non-differentiable product operations and explicit storage of per-frame posteriors, which again faces a **memory bottleneck** in high-dimensional radar sequences. We therefore adopt a discriminative classification framework to redesign the engineering implementation of TCA, transforming trajectory consistency prediction into an **auxiliary classification task**, *i.e.*, Track Loss.

Log-Space Temporal Aggregation. Inspired by the "product form" under the MAP perspective, we accumulate multi-frame semantic alignment signals in log space, preserving the mathematical intuition of probability products while obtaining numerical stability and computational efficiency:

$$\text{MAP-inspired Agg.: } \mathbf{z}_{\text{agg}} = \sum_{t=1}^{\tau} \log\Big(\phi(\mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger; \boldsymbol{\omega}_{\text{TCA}})\Big) + \mathbf{b} \quad \in \mathbb{R}^J, \text{where } \hat{\mathbf{x}}_t^\dagger \in \hat{\mathbf{H}}. \tag{3}$$

Here $\phi(\cdot,\cdot;\boldsymbol{\omega}_{\text{TCA}}) : \mathbb{R}^{2C} \to \mathbb{R}^J$ is composed of a learnable kernel function that operates in each candidate target-history pair processed by SAF, *i.e.*, $(\mathbf{x}^\star,\ \hat{\mathbf{x}}_t^\dagger)$, followed by a `Sigmoid` activation, used to estimate SAP; $\mathbf{b} \in \mathbb{R}^J$ is a learnable trajectory prior bias, corresponding to $\log P(\mathcal{T}^\star \mid \mathbf{x}^\star)$, *i.e.*, the class-dependent trajectory prior. Notably, no marginalization over classes is performed here, ensuring that multiple class-wise trajectory channels can be maintained simultaneously without explicitly branching the backbone to obtain posterior estimates; the logarithmic operation guaranties numerical stability, preventing underflow from multi-frame probability products.

Furthermore, to provide a certain "fault tolerance" in the temporal aggregation process, analogous to the temperature mechanism in SAF, we introduce a temporal temperature factor σ_{temp} to control the "strictness" of the trajectory consistency constraint. Specifically, prior to temporal aggregation, the per-frame class-wise SAP estimates are modulated via `SoftMax`:

$$\mathbf{m}_t = \frac{\exp(\mathbf{z}_t/\sigma_{\text{temp}})}{\sum_{i=1}^{\tau}\exp(\mathbf{z}_i/\sigma_{\text{temp}})}, \quad \text{where } \mathbf{z}_t = \log\Big(\phi(\mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger; \boldsymbol{\omega}_{\text{TCA}})\Big) \tag{4}$$

where σ_{temp} is the temporal temperature factor, controlling the "strictness" of trajectory consistency judgment. $\sigma_{\text{temp}} \to 0^+$: strict mode, requiring high consistency across most historical frames to yield a high probability; $\sigma_{\text{temp}} \to 1$: relaxed mode, allowing semantic drift in individual frames.

Accordingly, we obtain the modulated aggregation:

$$\boxed{\text{Modulated Agg.: } \mathbf{z}_{\text{agg}} = \sum_{t=1}^{\tau} \mathbf{m}_t \odot \log\left(\phi(\mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger; \boldsymbol{\omega}_{\text{TCA}})\right) + \mathbf{b} \quad \in \mathbb{R}^J} \tag{5}$$

Finally, the TCP is obtained via $\texttt{SoftMax}$ over $\mathbf{z}_{\text{agg}}$:

$$\text{TCP estimation: } \hat{\mathbf{p}} = \texttt{SoftMax}_j\left(\mathbf{z}_{\text{agg}}\right) \quad \in \Delta^J \tag{6}$$

OR-Aggregated Trajectory Label. To supervise the trajectory consistency task without storing intermediate gradients for all τ frames, we introduce a simple trajectory label construction using masks, $\{\mathbf{Y}_t\}_{t=1}^{\tau}$. Given a spatial location (h, w), the trajectory label $\mathbf{T}^\star \in \{0,1\}^J$ is defined as:

$$\mathbf{T}_j^\star = \bigvee_{t=1}^{\tau} \mathbf{1}\left[\mathbf{Y}_t(h,w) = j\right]. \tag{7}$$

Where $\mathbf{T}_j^\star = 1$ if and only if class j appears at coordinate (h, w) in at least one historical frame. This label serves as a sufficient statistic for "trajectory existence". That is, under the Markov assumption and local motion coherence assumption, if a target passes through this location within τ frames, its category must be reflected in the $\texttt{OR}$ aggregation, $\bigvee$. Then, τ frames of temporal information are compressed into a single J-dimensional vector, reducing memory complexity from $\mathcal{O}(B\tau HWJ)$ to $\mathcal{O}(BHWJ)$, which is a cost-benefit engineering design oriented toward deployment.

Track Loss. We treat trajectory consistency prediction as an auxiliary classification task running in parallel with the main segmentation task, employing the standard cross-entropy loss:

$$\boxed{\text{Track Loss: } \mathcal{L}_{\text{Trk}} = -\sum_{j=1}^{J} \mathbf{T}_j^\star \cdot \log \hat{p}_j}. \tag{8}$$

Since semantic segmentation assigns a single label per pixel, $\mathbf{T}^\star$ is typically one-hot or a sparse multi-hot vector. Track Loss compels the model to identify which semantic categories persist across the temporal window, thereby implicitly learning the temporal coherence of moving objects.

Finally, the total loss for TFoC can be written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Seg}} + \lambda_{\text{Trk}} \cdot \mathcal{L}_{\text{Trk}} \tag{9}$$

where $\mathcal{L}_{\text{Seg}}$ is the standard cross-entropy loss of the main segmentation head on $\mathbf{Y}_{\tau+1}$, and λ_{Trk} balances the strength of auxiliary supervision. During early training, λ_{Trk} can be set to a smaller value to prevent the auxiliary task from interfering with the main task; as convergence approaches, it can be increased to strengthen the temporal consistency constraint.

4 Experiments

To systematically and comprehensively validate the effectiveness of TFoC, we designed multiple sets of experiments, including ablation studies of SAF and TCA, an evaluation of TFoC under different input sequence lengths (the number of CPIs), and an overall comparison with other current state-of-the-art models. Meanwhile, to maintain generality, we adopted three mainstream public RSS datasets, covering an in-vehicle low-cost millimeter-wave continuous-wave radar and Ku-band low-altitude and maritime surveillance radars. These datasets were selected for multiple considerations, such as the temporal consistency across CPIs (the scanning mode in KuRALS-CW and the tracking mode KuRALS-PD), different spectral broadening of target signal (e.g., angular ambiguity in the RA view), and the accuracy of the annotations (e.g., the RA-view annotation correction in CARRADA-RAC). Please refer to Appendix B.1 and Appendix D.1 for detailed descriptions of the datasets. In addition, to ensure fairness and reproducibility, unless otherwise specified, we use a unified set of hyperparameters for all models. See B.2 for details of the training setup.

4.1 Effectiveness Analysis of SAF

SAF addresses the problem of how to identify relevant historical evidence to collect according to targets within a given spatial search range, and how to compress such evidence. It involves two

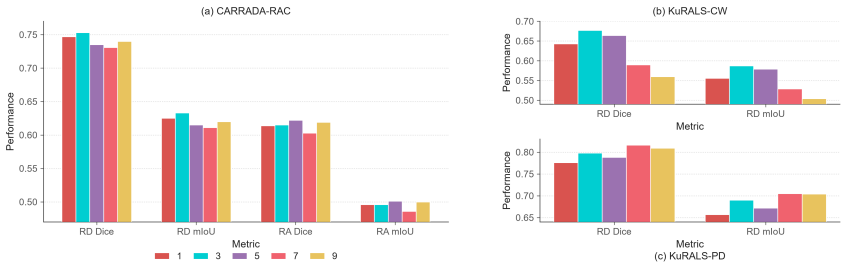


Figure 2: Performance trends under different `search_size`.

hyperparameters: the search size and the spatial temperature factor, denoted as N and σ_{spa} in Eq. 2. The former defines the evaluation scope for target-historical-feature-pair correlation (*i.e.*, SAP in Eq. 1), while the latter governs TFoC's confidence in SAP evaluations during spatial aggregation, or equivalently, the selection hardness for historical candidates. Therefore, we conducted experimental analysis on these two factors to investigate the impact of the SAF mechanism on TFoC performance.

Search Size. To simplify the formal formulation, we use N to denote a 2D search range in methodology. For the sake of generality, the one-sided notation k is used in experiments, where $k^2 = N$, which determines the scope within which SAF performs semantic alignment matching. The impact of k are reported in Fig. 2. The results show that the selection of the k requires comprehensive consideration of factors including the **temporal consistency** and the extent of **spectral broadening**. On CARRADA-RAC, TFoC performs best in RD at $k = 3$ and in RA at a moderately larger $k = 5$. This is because RD requires tight Doppler alignment, while RA needs a slightly wider search range to accommodate angular ambiguity. On KuRALS-CW, the performance decreases monotonically as k increases from 3 to 9. As this dataset own poor temporal consistency, a compact search window is sufficient for history-to-target alignment, whereas larger windows introduce clutter-induced mismatches. On KuRALS-PD, the trend is reversed, which indicates that its RD data benefits from a moderately enlarged search radius. However, what cannot be overlooked is that, compared with case $k = 1$ (w/o SAF), the target-history focus generally yields better results in most cases.

The Hardness of Spatial Focusing. In TFoC, the hardness of SAF, which is controlled by σ_{spa} in Eq. 2, governs how rigorously target frame units filter semantically consistent information across different historical frames. In Tab. 1, TFoC w/ `SoftMax` denotes the target frame and historical frames are jointly processed purely via `SoftMax`-modulated weights (most-relaxed SAF). TFoC w/ hard-selection denotes only the historical-frame points with top-1 SAF weight are retained (the strictest constraint). Results demonstrates that neither hard selection under maximum-a-posteriori estimation (Eq. 16) nor weight-softening treatments analogous to the attention mechanism (`SoftMax` case) appear to yield optimal RSS performance. This finding is critical for future extensibility-oriented research: for numerically sensitive data such as radar signals, we require both effective priors as constraints and a reasonable amount of learning latitude. Fig. 3-(a) explores the performance impact of fine-grained σ_{spa} settings. As targets appear as extended RD/RA units and cross-frame correspondences are typically neighborhood-level rather than unique, RD, RA, and overall metrics all rise with σ_{spa} and peak at 0.9. Specifically, an overly small σ_{spa} sharpens the semantic relation into near one-hot weights and risks locking onto false peaks (clutter and sidelobes, etc.), whereas $\sigma_{\text{spa}} \approx 0.9$ performs soft-weighted aggregation for multiple high-score spatial candidates that better matches target support and dilutes spurious peaks. Thus, the bottleneck of SAF is not insufficient sharpness but hard selection under non-unique radar correspondences. A moderately softened σ_{spa} better handles this uncertainty.

Table 1: Performance of different spatiotemporal focusing approaches.

Method	#Param.	RD-view (%)		RA-view (%)		Global
		mIoU	mDice	mIoU	mDice	Avg.
TFoC w/ `SoftMax`	5.1M	60.9	73.0	49.4	61.4	61.2
TFoC w/ hard-selection	5.1M	60.8	72.9	48.7	60.2	60.7
TFoC (default)	5.1M	**64.2**	**76.2**	**51.0**	**63.4**	**63.7**

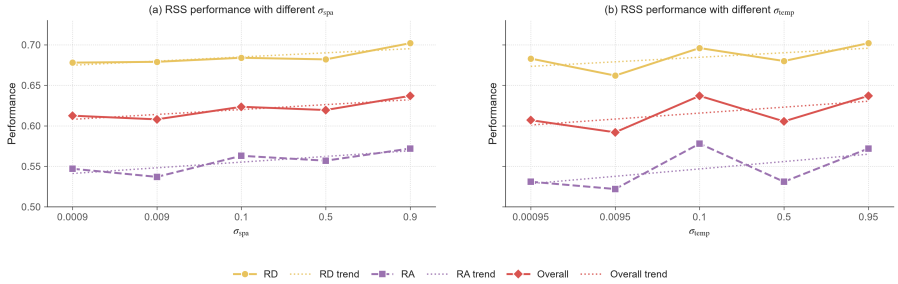


Figure 3: Average of mIoU and mDice under different σ_{spa} and σ_{temp} on CARRADA-RAC.

4.2 Effectiveness Analysis of TCA

Following the analysis of SAF, we proceed in this section to investigate the TCA within TFoC. As described in Sec. 3.2.2, to circumvent optimization difficulties arising from the non-differentiable property and memory bottleneck caused by explicit frame-wise TCP estimation via MAP, we reformulate it as a discriminative optimization problem parallel to segmentation. Specifically, we introduce a hyperparameter σ_{temp} (Eq. 4) in SAF into the forward-computation to govern the strictness of temporal aggregation. Then, an auxiliary loss function, Track Loss (Eq. 8), is incorporated to provide computationally affordable error-propagation grounds for temporal-consistency constraints.

The Strictness of Trajectory Consistency Judgment. Fig. 3-(b) explores the effect of different strictness of TCA, which controlled by σ_{temp} in Eq. 4. The overall metrics generally improve as σ_{temp} increases, yet the curves are non-monotonic. In RD, the evolution of the range-Doppler is comparatively continuous, so a longer path memory ($\sigma_{\text{temp}} \approx 0.95$) helps to accumulate evidence along true trajectories and suppress flickering clutter. However, the peak occurs at 0.1 in RA, coarse angular resolution and larger lateral displacements make long uniform trails prone to angular smearing and sidelobe mixing. Thus, the setting of σ_{temp} should be matched to the distinct trackability of RD and RA, rather than maximized indiscriminately.

Ablation Study of Track Loss. Track Loss serves as the explicit constraint for TCA. Theoretically, it can judge whether the aggregation output of TFoC conforms to the definition of "focusing". On CARRADA, removing Track Loss significantly degrades RA-view (-2.5% mIoU / -3.5% mDice). CARRADA-RAC further refines the RA annotations of CARRADA and the removal of Track Loss leads to a slight reduction in the degradation impact upon RA-view (-1.3% mIoU / -1.7% mDice). Track Loss delivers the largest gain on KuRALS-PD (75.1% vs. 73.2% without it, +1.9 points). However, TFoC without Track Loss performs best (61.7% vs. 61.0%) on KuRALS-CW. More details shown in Fig. 4. Overall, the benefit of Track Loss constraints is governed by the temporal consistency of radar data. Trajectory consistency supervision pays off when targets are continuously observed (CARRADA, CARRADA-RAC and KuRALS-PD), whereas it should be relaxed for data (like KuRALS-CW) with intermittent target revisits.[1].

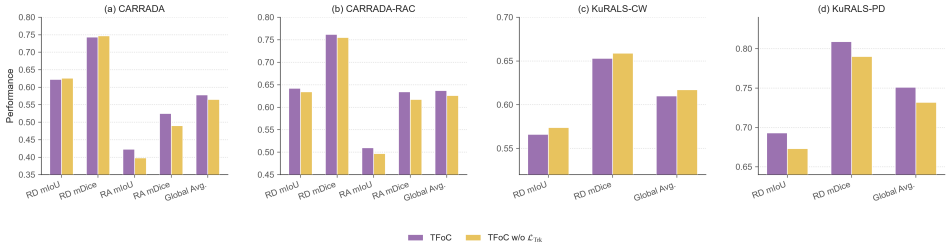


Figure 4: Ablation study of Track Loss.

4.3 Comparison with State-of-The-Art (SoTA)

Performance comparisons for different datasets are shown from Table 2 to Table 3. In the **CARRADA** group of Table 2, TFoC achieves a comparable average RSS performance (57.8%) to TARSS-Net

[1]More analysis of temporal continuity of different datasets is explored in Appendix D.1

Table 2: Comparisons with SoTA RSS networks on CARRADA and CARRADA-RAC.

Dataset	Method	#Param.	RD-view (%)		RA-view (%)		Global
			mIoU	mDice	mIoU	mDice	Avg.
CARRADA	**FCNLong et al. (2015)**	134.3M	54.7	66.3	34.5	40.9	49.1
	U-NetRonneberger et al. (2015)	17.3M	55.4	68.0	32.8	38.2	48.6
	DeepLabv3+Chen et al. (2017)	59.3M	50.8	61.6	32.7	38.3	45.9
	RSS-NetKaul et al. (2020)	10.1 M	32.1	36.9	32.1	37.8	34.7
	RAMP-CNNGao et al. (2021)	106.4 M	56.6	68.5	27.9	30.5	45.9
	MV-NetOuaknine et al. (2021a)	**2.4 M**	29.0	32.8	26.8	28.5	29.3
	MVA-NetOuaknine et al. (2021a)	4.8 M	53.5	65.3	37.1	44.8	50.2
	TMVA-NetOuaknine et al. (2021a)	5.6 M	56.1	68.0	37.7	46.2	52.0
	TransRadarDalbah et al. (2024)	4.9M	57.2	69.1	39.9	49.5	53.9
	T-RODNetJiang et al. (2023)	162.0M	-	-	**43.5**	53.6	-
	TransRSSZou et al. (2023)	-	60.4	73.0	43.0	**53.8**	57.6
	PKCIn-NetZhang et al. (2023b)	6.3M	60.7	72.6	43.1	53.7	57.5
	TARSS-Net_SZhang et al. (2024)	6.2 M	62.1	73.8	41.6	51.2	57.2
	TARSS-Net_DZhang et al. (2024)	6.3 M	**63.4**	**75.2**	41.4	51.3	**57.8**
	TFoC (ours)	5.1M	62.2	74.3	42.3	52.5	**57.8**
CARRADA-RAC	**TMVA-NetOuaknine et al. (2021a)**	5.6M	59.7	69.9	46.6	57.9	58.5
	PKCIn-NetZhang et al. (2023b)	6.3M	60.6	72.4	47.3	58.7	59.8
	TARSS-Net_SZhang et al. (2024)	6.2M	62.5	74.3	45.8	56.1	59.7
	TARSS-Net_DZhang et al. (2024)	6.3M	62.8	74.6	47.4	58.7	60.9
	TFoC (ours)	**5.1M**	**64.2**	**76.2**	**51.0**	**63.4**	**63.7**

Table 3: Performance comparison on KuRALS (CW: scanning mode; PD: tracking mode).

Method	#Param.	KuRALS-CW (%)			KuRALS-PD (%)		
		mIoU	mDice	Global avg.	mIoU	mDice	Global avg.
FCNLong et al. (2015)	134.3M	33.3	39.3	36.3	31.3	39.3	35.3
U-Net Ronneberger et al. (2015)	17.3M	47.2	56.5	51.9	46.6	58.7	52.7
DeepLabv3+Chen et al. (2017)	58.8M	50.5	58.8	54.7	53.3	66.3	59.8
HRNetSun et al. (2019)	65.8M	51.3	59.0	55.2	54.3	66.9	60.6
Swin-T Liu et al. (2021)	59.8M	51.9	59.1	55.5	52.5	63.7	58.1
SegFormer-B0 Xie et al. (2021)	3.7M	44.0	53.0	48.5	52.7	64.6	58.7
SegFormer-B1 Xie et al. (2021)	13.7M	48.8	59.2	54.0	52.6	65.0	58.8
RSSNetKaul et al. (2020)	10.1M	56.8	**67.5**	62.2	54.2	66.9	60.6
KuRALS-Net Li et al. (2026)	**1.2M/1.1M**	55.8	66.3	61.1	60.1	72.8	66.5
TMVA-Net^{su}Ouaknine et al. (2021a)	**1.2M**	-	-	-	52.9	63.1	58.0
PKCIn-Net^{su}Zhang et al. (2023b)	**1.2M**	-	-	-	56.7	65.9	61.3
TARSS-Net_S^{su}Zhang et al. (2024)	**1.2M**	-	-	-	53.2	63.8	58.5
TARSS-Net_D^{su}Zhang et al. (2024)	1.3M	-	-	-	58.4	67.1	62.8
TFoC^{su} (ours)	1.8M	**58.4**	66.8	**62.6**	**77.1**	**85.7**	**81.4**

with fewer parameters and remains advantageous. It should be noted that TFoC uses a search size of 5 for both RD and RA views, whereas the frequency extent of the targets differs between views, which may explain the performance gap between TFoC and TARSS-Net_D from different points of view. **CARRADA-RAC** provides a more accurate standardized correction for RA annotations, which improves the performance of all methods on the RA view. With more precise labels, TFoC can perform spatiotemporal focusing more accurately. Consequently, its RA performance increases by approximately 9%–11% compared to that on CARRADA, achieving superior performance over existing state-of-the-art methods. **KuRALS** provides only single-view RD data. As shown in Table 3, with the continuous-wave radar system (KuRALS-CW), the long interval between data acquisitions leads to weak temporal consistency, preventing TFoC from fully leveraging its temporal focusing capability. Its advantage over existing models is more evident on the pulse-doppler rada data (KuRALS-PD), i.e., data with stronger characteristics of temporal consistency, where its performance can exceed the current best model by more than 15%.

5 Conclusion

Summary: inspired by physical intuition of camera gimbal, TFoC achieves a novel spatio-temporal alignment mechanism, *i.e.* feature gimbal. In methodology discussion, we start from maximum-a-posterior estimation and Bayesian theory, and further perform approximate solutions considering the requirements of engineering implementation, which provides a theoretical basis for realizing the physical intuition behind TFoC. The experimental section has comprehensively verified the effectiveness of TFoC from multiple perspectives. **Limitations**: there are no fixed rules for the selection of some critical hyperparameters, which requires preliminary estimation based on the consistency of scene data. **Future work**: this model supports scaling in temporal processing scope, laying a foundation for the subsequent construction of long-term memory for radar large models.

STATEMENTS

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, *i.e.*, generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims or assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results; all core methods, experiments, and conclusions of this paper are originally proposed and conducted by the authors. Additionally, we used generative AI tools for tasks with recommended disclosure, *i.e.*, editing this paper to improve readability and assisting with format adjustment. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only public radar-collected dataset; no human subjects, personally identifiable information, or ethically sensitive content are involved, and neither crowdsourcing nor IRB-related procedures apply. All datasets introduced in this paper, *i.e.*, CARRADAR, CARRADA-RAC and KuRAls are properly cited and used in compliance with their open-source licenses. We also recognize the dual-use nature of radar perception technology: while TFoC benefits civilian domains such as autonomous navigation safety, airspace monitoring, low-slow-small UAV management, and disaster search and rescue, it could be misused for unauthorized surveillance or military offensive purposes. To mitigate this risk, we will append a usage restriction statement to the code repository license and to a prominent position in the paper upon acceptance, explicitly limiting the model to applications serving positive societal purposes only. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have submitted the project code together with the model containing typical parameters as a supplementary attachment.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824, 2021.

Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for lidar sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 760–767, 2018. doi: 10.1109/IVS.2018.8500543.

Daniel Brodeski, Igal Bilik, and Raja Giryes. Deep radar detector. In *2019 IEEE Radar Conference (RadarConf)*, pp. 1–6, 2019. doi: 10.1109/RADAR.2019.8835792.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6299–6308, 2017.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. URL `http://arxiv.org/abs/1706.05587`.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for

statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. pp. 4846–4855, 10 2017. doi: 10.1109/ICCV.2017.518.

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Yahia Dalbah, Jean Lahoud, and Hisham Cholakkal. Transradar: Adaptive-directional transformer for real-time multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 353–362, 2024.

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Amir Karami, Mohammad Arzani, Luc Van Gool, and Rahman Yousefzadeh. Temporal 3d convnets: New architecture and transfer learning for video classification. 11 2017.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2758–2766, 2015. doi: 10.1109/ICCV.2015.316.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: 10.1207/s15516709cog1402_1.

Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal*, 21(4):5119–5132, 2021. doi: 10.1109/JSEN.2020.3036047.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Shun-Hao Jiang, Xian Zhang, Vishal Monga, Srinivas Venkatesh, and Jarvis Haupt. T-rodnet: Transformer for vehicular millimeter-wave radar object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023. doi: 10.1109/TIM.2023.3246773.

Prannay Kaul, Daniele de Martini, Matthew Gadd, and Paul Newman. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 431–436, 2020. doi: 10.1109/IV47402.2020.9304674.

Teng Li, Liwen Zhang, Youcheng Zhang, Zijun Hu, Pengcheng Pi, Zongqing Lu, Qingmin Liao, and Zhe Ma. AdaPKC: PeakConv with adaptive peak receptive field for radar semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136545–136575, 2024.

Teng Li, Qingmin Liao, Youcheng Zhang, Xinyan Zhang, Zongqing Lu, and Liwen Zhang. Kurals: Ku-band radar datasets for multi-scene long-range surveillance with baselines and loss design. *Remote Sensing*, 18(1):173, 2026. doi: 10.3390/rs18010173.

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 (ICCV)*, pp. 10012–10022, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

B. Major, D. Fontijne, A. Ansari, R. T. Sukhavasi, and S. Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15671–15680, October 2021a.

Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range- angle- doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5068–5075, 2021b. doi: 10.1109/ICPR48806.2021.9413181.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

M. A. Richards. *Fundamentals of Radar Signal Processing, Second Edition*. Fundamentals of Radar Signal Processing, Second Edition, 2005.

Hermann Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983. doi: 10.1109/TAES.1983.309350.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, 2019. doi: 10.1109/CVPR.2019.00584.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419, 2020.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):954–967, 2021. doi: 10.1109/JSTSP.2021.3058895.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7444–7452, 2018.

Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, Yanwei Yu, and Zhenhui Jessie Li. Modeling spatial-temporal dynamics for traffic prediction. *ArXiv*, abs/1803.01254, 2018. URL `https://api.semanticscholar.org/CorpusID:3698767`.

F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

Liwen Zhang, Jian Pan, Youcheng Zhang, Yuanpei Chen, Zhe Ma, Xuhui Huang, and Kewu Sun. Capturing temporal-dependence in radar echo for spatial-temporal sparse target detection. *Journal of Radars*, 12:356, 2023a. ISSN ISSN 2095-283X (Print). doi: 10.12000/JR22228. URL https://radars.ac.cn/article/doi/10.12000/JR22228.

Liwen Zhang, Xinyan Zhang, Youcheng Zhang, Yufei Guo, Yuanpei Chen, Xuhui Huang, and Zhe Ma. Peakconv: Learning peak receptive field for radar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17577–17586, 2023b.

Youcheng Zhang, Liwen Zhang, Zijun Hu, Pengcheng Pi, Teng Li, Yuanpei Chen, Shi Peng, and Zhe Ma. Tarss-net: Temporal-aware radar semantic segmentation network. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 4906–4933. Curran Associates, Inc., 2024. doi: 10.52202/079017-0159. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/08a362bd4ae1934e099ce025f06039fe-Paper-Conference.pdf.

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 (CVPR)*, June 2019.

Hao Zou, Zhen Xie, Jiarong Ou, and Yutao Gao. Transrss: Transformer-based radar semantic segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6965–6972, 2023. doi: 10.1109/ICRA48891.2023.10161200.

A Theoretical Interpretation of TFoC

This appendix provides the complete probabilistic interpretation of TFoC summarized in Sec. 3.1 and Sec. 3.2 of the main text. Unless otherwise specified, all notations follow Table 4.

Table 4: Notation table for TFoC.

Name	Notation	Description
Input	$\mathbf{X}^0 \in \mathbb{R}^{(\tau+1)\times R\times D}$	τ, R, D: historical frames, range bins, Doppler bins.
LSC	$\Phi(\cdot;\boldsymbol{\omega}_{\text{LSC}}):\mathbb{R}^{(\tau+1)\times R\times D}\to\mathbb{R}^{C\times(\tau+1)\times H\times W}$	Feature extraction front-end (Conv or PeakConv). C, H, W: output channels, compressed range/Doppler dims.
LSE	$\mathbf{X}\in\mathbb{R}^{C\times(\tau+1)\times H\times W}$	Output of LSC, input of TFoC.
Masks	$\mathbf{Y}=\{\mathbf{Y}_t\}_{t=1}^{\tau+1}\in\mathbb{B}^{J\times(\tau+1)\times R\times D}$	GT masks with J semantic classes for input RD matrices.
Backbone	$\texttt{Backbone}(\cdot;\boldsymbol{\Omega}):\mathbb{R}^{(\tau+1)\times R\times D}\to\mathbb{R}^{J\times R\times D}$	Normal RSS task; only $\mathbf{Y}_{\tau+1}$ is used for supervision.
Historic LSE	$\{\mathbf{X}_t^\dagger\}_{t=1}^{\tau}\doteq\{\mathbf{X}_t\}_{t=1}^{\tau}$	$\dagger$ denotes historical information.
Target LSE	$\mathbf{X}^\star\doteq\mathbf{X}_{\tau+1}$	$\star$ denotes the target/current frame.
HSF at t	$\mathcal{H}_t=\{\mathbf{x}_{t,n}^\dagger\}_{n=1}^{N}\in\mathbb{R}^{C\times N}$	Search field for target unit $\mathbf{x}^\star\in\mathbb{R}^C$ at t-th historic LSE; search size $N=s^2$.
Global HSF	$\mathcal{H}=\{\mathcal{H}_t\}_{t=1}^{\tau}$	HSF across all historic frames.
Class index	$j=1,\dots,J$	$j^\dagger, j^\star$: class labels for historic/target feature points.

A.1 A Maximum A Posteriori Perspective on SAF

Assumption. Moving objects exhibit coherent spatial variation and semantic consistency over a local spatio-temporal neighborhood. Within such a local window, semantic alignment can be approximated as the identity of the moving object itself.

Problem Formulation. Given a target feature point $\mathbf{x}^\star \in \mathbf{X}^\star$ from the current-frame LSE, and a candidate historical feature point $\mathbf{x}^\dagger_{t,n} \in \mathcal{H}_t$ from some historical timestep t, our goal is to estimate the posterior probability that the two share the same semantic label and to use this as a criterion to select, from each HSF $\mathcal{H}_t$, the candidate most semantically similar to $\mathbf{x}^\star$. With high probability (taken as 1 in this paper), this selected point is the representation of the same physical entity observed at timestep t. We term this posterior the **Semantic Alignment Probability (SAP)**:

$$\boxed{\text{SAP}: \; P\left(j^\star = j^\dagger_{t,n} \mid \mathbf{x}^\star, \; \mathbf{x}^\dagger_{t,n}\right).} \tag{10}$$

Under the *Conditional Independence Assumption (CIA)*, *i.e.*, the semantic class distributions are independent given either the target or historical LSE, we further decompose the SAP as the product of two conditional probabilities as follows:

$$P\left(j^\star = j^\dagger_{t,n} \mid \mathbf{x}^\star, \mathbf{x}^\dagger_{t,n}\right) = \sum_{j=1}^{J} P\left(j^\star = j \mid \mathbf{x}^\star\right) \cdot P\left(j^\dagger_{t,n} = j \mid \mathbf{x}^\dagger_{t,n}\right). \tag{11}$$

This formulation transforms the SAP into a marginalization over classes of two posterior probabilities. Both terms can be estimated directly by the RSS backbone. Letting $\mathbf{q}^\star, \mathbf{q}^\dagger_{t,n} \in \Delta^J$ denote the `SoftMax` output of the RSS backbone:

$$\mathbf{q}^\star = \texttt{SoftMax}\left(\texttt{Backbone}(\mathbf{x}^\star;\, \mathbf{\Omega})\right), \tag{12}$$

$$\mathbf{q}^\dagger_{t,n} = \texttt{SoftMax}\left(\texttt{Backbone}(\mathbf{x}^\dagger_{t,n};\, \mathbf{\Omega})\right), \tag{13}$$

we compute the SAP per-class (which serves as the semantic alignment score between $\mathbf{x}^\star$ and $\mathbf{x}^\dagger_{t,n}$):

$$\alpha_{t,n,j} = q^\star_j \cdot q^\dagger_{t,n,j}. \tag{14}$$

Then we can have the class-agnostic marginal alignment probability:

$$\mathcal{A}_{t,n} = \sum_{j=1}^{J} \alpha_{t,n,j} = \mathbf{q}^{\star\top} \mathbf{q}^\dagger_{t,n} \in [0, 1]. \tag{15}$$

Maximum A Posteriori for Hard Focusing. Ideally, for each historical timestep t, we can perform MAP estimation to purify the information contained in $\mathcal{H}_t$:

$$\hat{n}_t \triangleq \arg\max_n \left\{\mathcal{A}_{t,n}\right\}_{n=1}^{N} = \arg\max_n \left\{ \sum_{j=1}^{J} q^\star_j \cdot q^\dagger_{t,n,j} \right\}_{n=1}^{N}. \tag{16}$$

Subsequently, we collect the complete historically focused feature sequence for each $\mathbf{x}^\star$:

$$\hat{\mathcal{H}} = \left\{\hat{\mathbf{x}}^\dagger_t\right\}_{t=1}^{\tau} = \left\{\mathbf{x}^\dagger_{t,\hat{n}_t}\right\}_{t=1}^{\tau}. \tag{17}$$

MAP achieves focusing behavior through **hard selection**: each historical frame's HSF contributes exactly one candidate feature point that is most semantically coherent with $\mathbf{x}^\star$.

A.2 A Bayesian Perspective on Trajectory Consistency

Trajectory consistency. We define trajectory consistency as a binary event $\mathcal{T}^\star \in \{0, 1\}$:

$$\text{Trajectory consistency event: } \mathcal{T}^\star = 1 \Leftrightarrow \hat{\mathcal{H}} \text{ belongs to the same traj. as } \mathbf{x}^\star. \tag{18}$$

Objective. Given the definition of the event, the objective is to maximizing the trajectory consistency posterior (TCP) as follows:

$$\boxed{\text{TCP: } P(\mathcal{T}^\star = 1 \mid \hat{\mathcal{H}}, \mathbf{x}^\star) = \frac{\overbrace{P(\hat{\mathcal{H}} \mid \mathcal{T}^\star = 1, \mathbf{x}^\star)}^{\text{jointly cond. likelihood}} \cdot \overbrace{P(\mathcal{T}^\star = 1 \mid \mathbf{x}^\star)}^{\text{traj. prior}}}{\underbrace{P(\hat{\mathcal{H}} \mid \mathbf{x}^\star)}_{\text{marginal likelihood, independent of } \mathcal{T}^\star}}.} \tag{19}$$

For the denominator of TCP, it represents a marginalization over $\mathcal{T}^\star$ and is therefore independent of $\mathcal{T}^\star$ itself. During MAP optimization of TCP, it can be treated as a constant:

$$P(\hat{\mathcal{H}} \mid \mathbf{x}^\star) = \sum_{\mathcal{T}} P(\hat{\mathcal{H}} \mid \mathcal{T}, \mathbf{x}^\star) \cdot P(\mathcal{T} \mid \mathbf{x}^\star) \doteq \text{const}. \tag{20}$$

Markov Decomposition. Under the Markov assumption, where observations at successive timesteps are conditionally independent given a trajectory, we can decompose the jointly conditioned likelihood numerator as:

$$P(\hat{\mathcal{H}} \mid \mathcal{T}^\star = 1, \mathbf{x}^\star) = \prod_{t=1}^{\tau} P(\hat{\mathbf{x}}_t^\dagger \mid \mathcal{T}^\star = 1, \mathbf{x}^\star). \tag{21}$$

From Trajectory Consistency to Semantic Alignment. Evidently, if the event $\mathcal{T}^\star = 1$ occurs, then $\mathbf{x}^\star$ and $\hat{\mathbf{x}}_t^\dagger$ must be semantically aligned; that is, **semantic alignment is a necessary condition for trajectory consistency**. However, semantic alignment alone is insufficient to guaranty trajectory consistency; *e.g.*, two distinct moving objects may belong to the same semantic category. Therefore, we cannot directly equate SAP in Eq. 10 with trajectory likelihood, but we can use the former as a **discriminative surrogate** for the latter van den Oord et al. (2018).

Specifically, in the absence of an explicit motion model, **we assume** that given trajectory consistency, *the stronger the semantic alignment of historical observations, the more credible the trajectory hypothesis*. This leads us to approximate the joint conditional likelihood as the product of per-frame semantic alignment posteriors:

$$P(\hat{\mathcal{H}} \mid \mathcal{T}^\star = 1, \mathbf{x}^\star) \approx \prod_{t=1}^{\tau} P(j^\star = j_t^\dagger \mid \mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger). \tag{22}$$

where $P(j^\star = j_t^\dagger \mid \mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger)$ is precisely the SAP defined in SAF. Substituting the above into Eq. 19 and ignoring the marginal likelihood denominator independent of $\mathcal{T}^\star$ under the MAP framework, we obtain:

$$\max_{\mathcal{T}} P(\mathcal{T}^\star \mid \hat{\mathcal{H}}, \mathbf{x}^\star) \equiv \max_{\mathcal{T}} \prod_{t=1}^{\tau} P(j^\star = j_t^\dagger \mid \mathbf{x}^\star, \hat{\mathbf{x}}_t^\dagger) \cdot P(\mathcal{T}^\star \mid \mathbf{x}^\star). \tag{23}$$

Here $P(\mathcal{T}^\star = 1 \mid \mathbf{x}^\star)$ denotes the prior probability that the target belongs to some motion trajectory, conditioned on the target feature. In practical implementation, we treat it as a learnable class-dependent bias, reflecting that different categories possess distinct motion continuity priors.

Under the MAP perspective, maximizing the trajectory consistency posterior is equivalent to maximizing the joint of "semantic alignment probability $\times$ trajectory prior." This makes TCA simultaneously leverage the semantic alignment of SAF and additional temporal supervisory signals.

B Experimental Details

B.1 Datasets

CARRADA Dataset. Ouaknine et al. (2021b) CARRADA is a large-scale camera–radar synchronized dataset. It collected under diverse scenes and weather conditions by FMCW radar with multi-view annotated recordings (RAD tensors) . There are four classes (*pedestrian*, *cyclist*, *vehicle*, *background*), which allows multiple objects per frame. The RAD tensors of size is $256 \times 256 \times 64$. We follow the train/val/test split in Ouaknine et al. (2021a).

CARRADA-RAC Dataset. Zhang et al. (2023b) CARRADA RD/RA annotations were generated semi-automatically by associating camera and radar data via Mean-Shift tracking Comaniciu & Meer (2002) in a unified physical domain (3D Cartesian + DoA + Doppler). Due to unreliable camera depth and the limited angular resolution of low-cost FMCW radar, the resulting clusters could deviate from true locations, especially in the RA view. CARRADA-RAC recalibrate the RA annotations and release a higher-quality variant.

KuRALS. Li et al. (2026) This dataset is collected by Ku-band radars and oriented to both maritime and land surveillance scenarios. It covers multiple target categories including unmanned aerial

vehicles (UAVs), ships, vehicles, and pedestrians. Equipped with various annotations such as semantic segmentation, it can support the training and validation of diverse radar intelligent models. KuRALS consists of two subsets: the continuous-wave radar subset KuRALS-CW and the pulse-Doppler radar subset KuRALS-PD, whose RD tensor dimensions are 2048×128 and 800×64 respectively. Since KuRALS contains only RD tensors, models marked with ^{sv} are adapted to a single-view variant when needed.

B.2 Training Setup

Hyperparameter: In the experiments presented in Sec. 4, the proposed TFoc model adopts default hyperparameters: spatial-focusing "hardness" control factor σ_{spa} is set to 0.9, and the temporal decay factor σ_{temp} is set to 0.95. The search window size and input time-series length of the model vary across different datasets. For the CARRADAR and CARRADAR-RAC datasets, the TFoc model takes 9 frames as input ($t = 9$) with a search window size of 5 ($k = 5$). On the KuRALS dataset, CW defaults $t = 7$ and $k = 5$, as well as PD defaults $t = 9$ with $k = 3$. In the SOTA comparison shown in Tab. 3, we adopted $t = 8$ with $k = 5$ for KuRALS-CW and $t = 12$ with $k = 3$ for the KuRALS-PD.The performance of all other models is consistent with the results reported in their respective published papers.

Evaluation metrics: All the models are evaluated with cell-wise precision/recall, Intersection over Union (IoU), and Dice score during training/validation phase. And the mean IoU (mIoU) and mean Dice (mDice) over categories are used as the principle metrics for model performance comparison. All the reported results were obtained from the test subset.

C Impact of Temporal Modeling Methods

In the default configuration, TFoC follows the principle of causal temporal modeling, taking the last temporal frame as the target frame. In this section, we further investigate the non-causal temporal modeling scheme, which selects the middle frame as the target frame, donated as TFoC (m-tgt.), and performs semantic consistency alignment and trajectory consistency constraints on the temporal frames before and after it.

Table 5: Impact of Temporal Modeling Methods.

Method	#Param.	RD-view (%)		RA-view (%)		Global
		mIoU	mDice	mIoU	mDice	Avg.
CARRADA						
TFoC	5.1M	62.2	74.3	**42.3**	**52.5**	**57.8**
TFoC (m-tgt)	5.1M	**63.1**	**75.1**	40.9	50.5	57.4
CARRADA-RAC						
TFoC	5.1M	**64.2**	**76.2**	**51.0**	**63.4**	**63.7**
TFoC (m-tgt)	5.1M	63.2	75.4	50.4	62.6	62.9
KuRALS-CW						
TFoC	1.8M	**56.6**	**65.3**	-	-	**61.0**
TFoC (m-tgt)	1.8M	55.3	63.5	-	-	59.4
KuRALS-PD						
TFoC	1.8M	69.3	80.9	-	-	75.1
TFoC (m-tgt)	1.8M	**69.5**	**81.0**	-	-	**75.3**

As shown in Tab. 5, there is no substantial performance gap between TFoC (mtgt) and TFoC across all four datasets. TFoC is built upon the principle of feature gating and excels at capturing semantically homogeneous information within time series. Therefore, whether for causal temporal modeling with the last frame as the anchor frame or non-causal temporal modeling anchored at intermediate frames, TFoC essentially performs semantic consistency measurement between units on target and non-target frames and increases the density of valid information after alignment. It can thus be further inferred that taking any frame as a continuous node (anchor frame) within consecutive time sequences barely affects the performance of TFoC.

D HANDLING CAPACITY FOR TIME LENGTH — TEMPORAL SCALING.

TFoC extracts temporal correlation information of targets from continuous data sequences. It can take inputs of arbitrary lengths (i.e., frame/CPI numbers) with fixed network parameters. The length of the time series fed into TFoC not only affects the model's performance in RSS tasks but also reflects its temporal scaling capability.

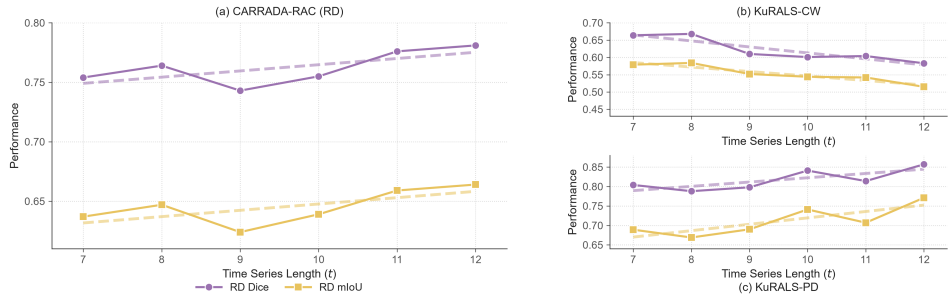


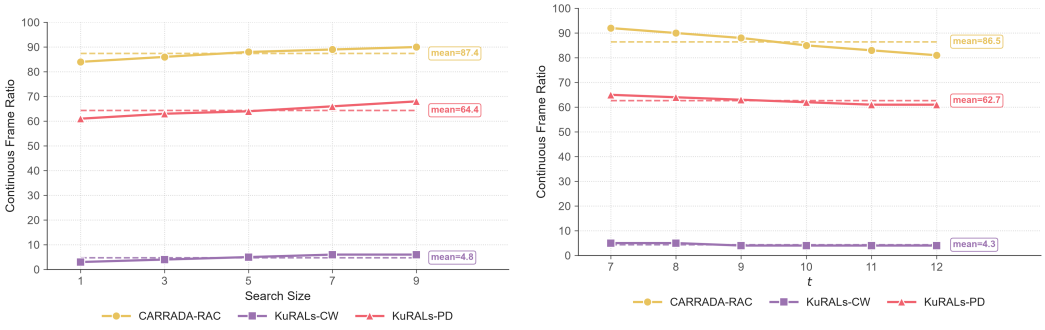
Figure 5: Performance trends under different frame numbers.

Figs. 5 evaluates the TFoC under varying input lengths t, where $t-1$ history frames are aggregated onto the last target frame through alignment of temporal-spatial consistency. On **CARRADA-RAC (RD)**, performance improves gradually and peaks around t=12 (Dice: 0.781; mIoU: 0.664). Similarly, **KuRALS-PD** also consistently benefits from longer sequences, reaching its best score at t=12 (Dice: 0.857; mIoU: 0.771). In these two datasets, TFoC shows that its overall performance improves with increasing t, revealing its ability to scale with the temporal sequence length. However, **KuRALS-CW** is highly sensitive: performance peaks at t=8 (Dice: 0.668) but drops sharply at t=9 (Dice: 0.610). As KuRALS-CW exhibits less temporal consistency than CARRADA-RAC and KuRALS-PD, too long windows introduce misaligned historical cues, and its overall performance degrades as t increases.

In general, for datasets with short temporal consistency, such as KuRALS-CW, the time-window length that yields the optimal performance is correspondingly shorter. On data sequences with long-range temporal consistency (CARRADA-RAC and KuRALS-PD), appropriately increasing the length of the time series to be aligned and focused helps optimize the RSS performance. **TFoC exhibits strong temporal scaling capability for datasets with high temporal continuity.**

D.1 TEMPORAL CONSISTENCY ANALYSIS OF THE DATASET

Due to differences in radar systems and data acquisition approaches, there are considerable variations in inter-frame continuity across different datasets. Analyzing the temporal continuity of targets within datasets helps to further validate several observations from the experiments in this paper. Fig. 6 illustrates the variation of Continuous Frame Ratio in different conditions.



(a) Continues frame rate under different search size.

(b) Continues frame rate under different temporal sequence lengths.

Figure 6: The exploration on Continues frame rate of different datasets.

The calculation for the Continuous Frame Ratio is defined as follows: for each window of search size, first computing the target-history frame-pair continuity rate as `rate = the number of`

`historical frames correlated with the target frame / (t-1)` (t is temporal sequence lengths input to TFoC) , then take the average over all windows. For a frame-pair consisting of the target frame T and historical frame $T-k$, continuity (*i.e.* `historical frames correlated with the target frame`) is defined when the per-axis gaps between the target block in the target frame and its corresponding target block of the same category in the historical frame are all $\leq (\text{search_size}-1)/2$ bins; otherwise, it is defined as a break point.

It can be seen from Fig. 6 (a), the temporal continuity of the CARRADA-RAC and KuRALs-PD datasets is evidently superior to that of the KuRALs-CW dataset. Continuous Frame Ratio increases as search size grows. However, this does not imply that the performance of TFoC will improve accordingly, since increasing search size also introduces more interferences such as clutter. This provides complementary evidence for our analysis in `Search Size` (Sec. 4.1): as KuRALS-CW owns poor temporal continuity, a compact search window is sufficient for history-to-target alignment, whereas larger windows expand the effective candidate set ($(\text{search_size})^2$) and introduce clutter-induced mismatches.

In Fig. 6 (b), the temporal continuity of the CARRADA-RAC and KuRALs-PD datasets remains substantially higher than that of the KuRALs-CW dataset. Therefore, in Sec. D, TFoC exhibits an opposite trend in the KuRALs-CW dataset compared to the other two datasets. On the other hand, Continuous Frame Ratio decreases as t grows, which is readily understandable given the calculation procedure of Continuous Frame Ratio mentioned above. But it still maintains a high degree of temporal continuity on CARRADA-RAC and KuRALs-PD. Furthermore, TFoC can automatically mine and align consistent temporal information through SAF, TCA and other operations on the input data sequences. Accordingly, in Sec. D, the performance of TFoC presents a steadily growing trend as t increases on dataset with high degree of temporal continuity, demonstrating its scalability potential for long-term sequential data.

Notably, the Continuous Frame Ratio of KuRALs-CW remains at an extremely low level, below 5%. This indicates that there is almost no inter-frame temporal continuity for KuRALs-CW. Nevertheless, TFoC does not distinguish between target and non-target information when performing spatial and temporal consistency alignment. That is to say, on the KuRALs-CW dataset, TFoC can also complete the RSS task by aligning the consistency of non-target information. In this case, forcibly imposing spatial-temporal consistency constraints on targets will instead degrade the model performance. This also explains the experimental phenomenon in `Ablation Study of Track Loss` (Sec. 4.2) where the performance of TFoC is improved after removing the track loss on the KuRALs-CW dataset while degraded of varying degrees on other datasets.

E VISUALIZATION

To provide an intuitive illustration of the procedures and effects of spatiotemporal focusing, this section visualizes the features produced at key steps. Using single-view RD data as an example, we first randomly select nine contiguous frames from the KuRALS-PD dataset and take the last frame as the target frame, as shown in Fig. 7. The corresponding ground-truth semantic segmentation masks are shown below the data accordingly.

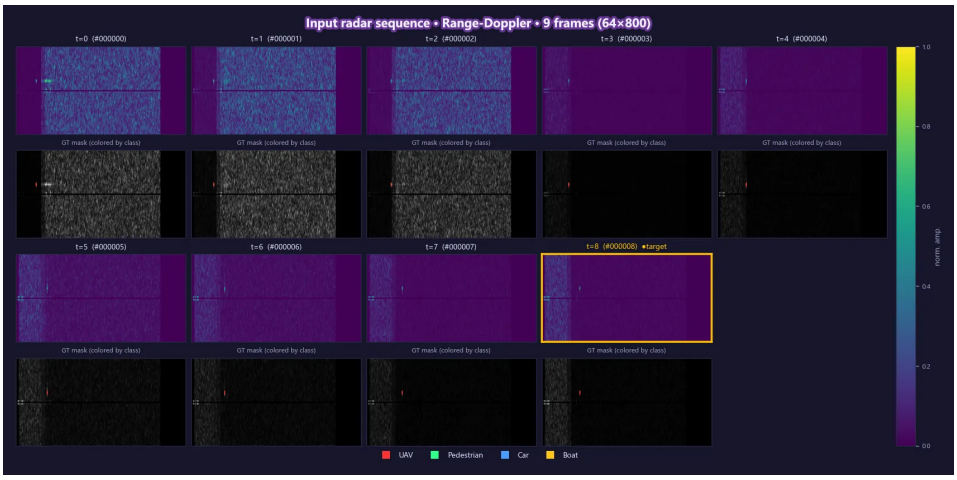


Figure 7: Input examples.

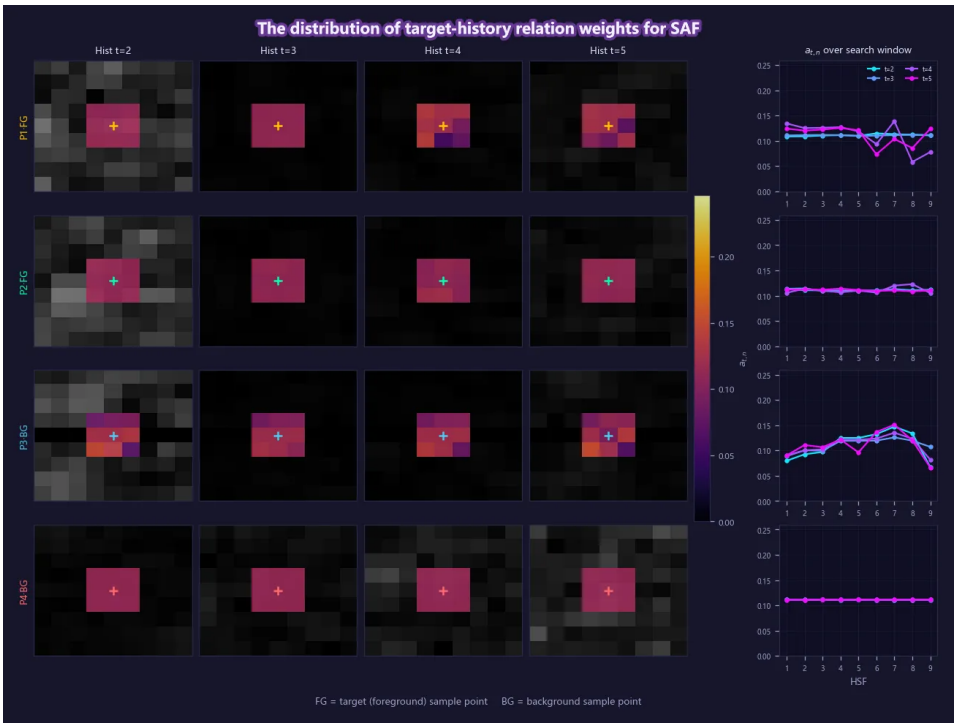


Figure 8: The distribution of target-history relation weights for SAF ($a_{t,n}$ in Eq. 2). Left: the spatial visualization of weights for 4 historical search fields (3×3) of a given anchor point; Right: the curve plot view of weight scores for 9 spatial positions of each 3×3 historical search field.

We randomly sample four anchor units from the target frame, including two target units and two background units. For each unit, the semantic-consistency measurement (SAF) is performed with the units within the search region (3×3 in the example) in the historical frames, and the measurement results are shown in Fig. 8. On the vertical axis, P1–P4 means four sampled points. ”FG” and ”BG” stand for foreground targets and background points, respectively. For example, ”P1 FG” means the first sampled foreground point on the target frame. Each row visualizes the measurement results, which are encoded by color. Strong color contrast in the search region implies large semantic disparities and the brightest unit has the highest semantic similarity to the anchor unit (like P1 FG, t=4; P3 BG, t=2). Uniform color distribution means units in search region are either semantically close to or far from the anchor unit (like P2 FG; P4 BG). In the right column of the figure, the nine measurement values within the search field are unfolded and plotted as a line chart, thereby showing the variation trend.

We further map the different semantic-measurement results back to their positions in the historical frames, as shown in Fig. 9. Then, the semantic-measurement results within the search region are filtered or fused in different hard ways, thereby computing an SAF result for each unit under test. Fig. 10 visualizes the resulting feature tensors of each historical frame after SAF (*i.e.* Eq.2). It can be seen that, by aligning the target–history semantic relations for each query cell on the target frame, the feature elements on each historical frame are indeed rearranged (compared with the historical-frame background image in Fig. 9) , and distinct pattern characteristics are exhibiting.

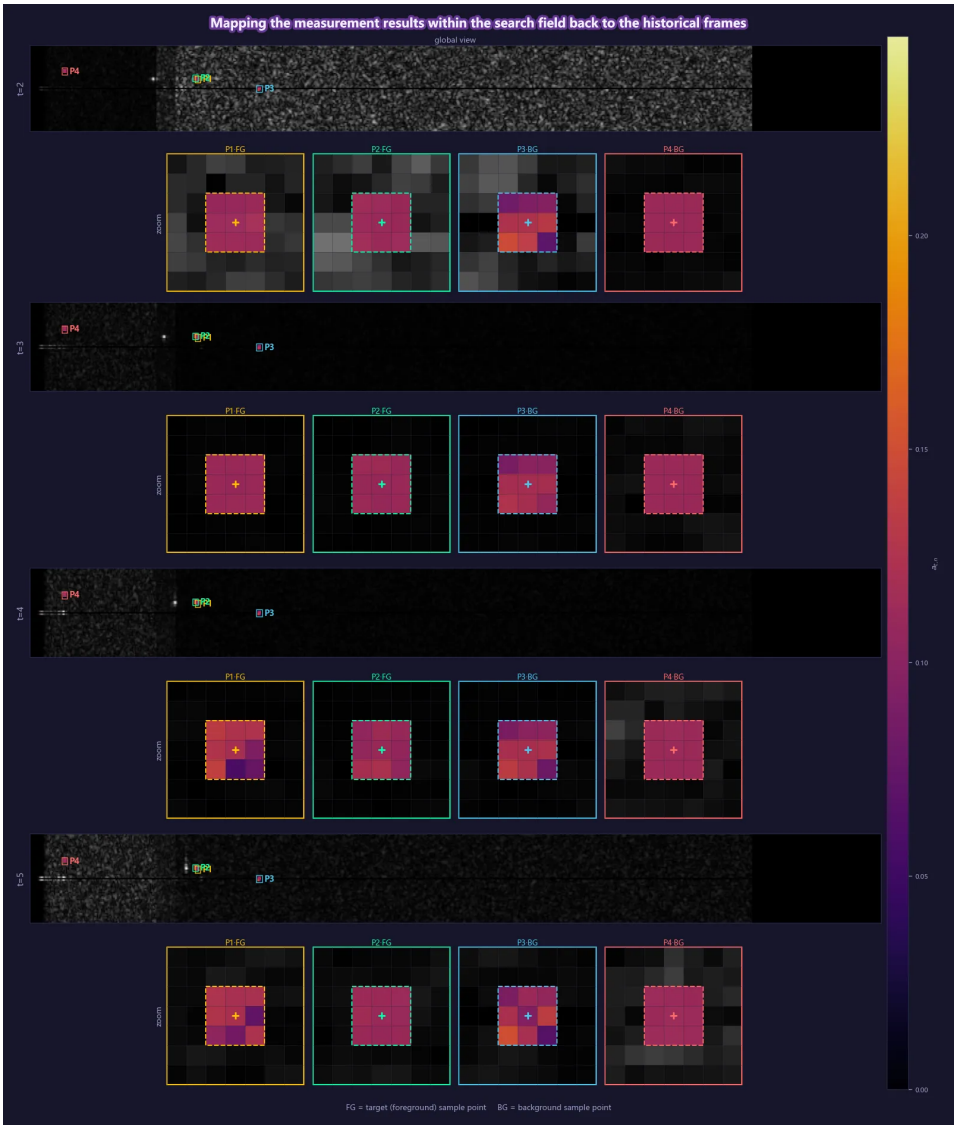


Figure 9: Mapping the measurement results within the search field back to the historical frames.



Figure 10: Spatial focusing results on each historical frame (the results of Eq.2).

In SAF, we achieve local spatial focusing via the target–history semantic relation measurement. Trajectory Consistency-based Temporal Aggregation (TCA) further provides TFoC with a temporal-consistency constraint, completing its full temporal focusing capability.

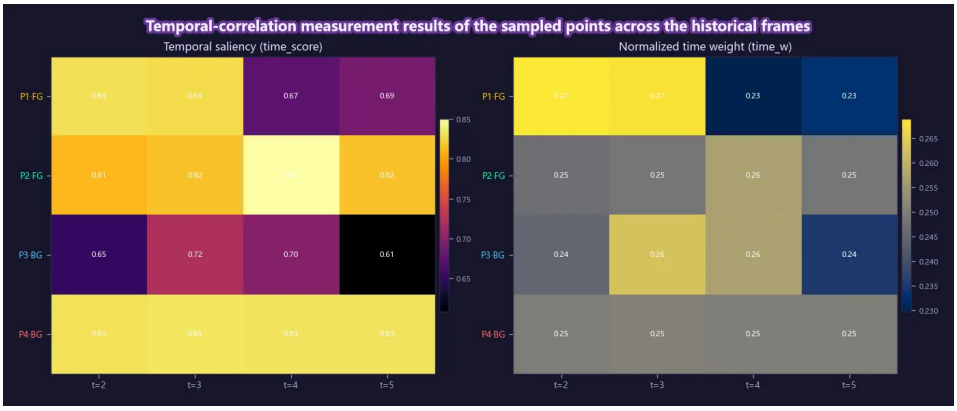


Figure 11: Temporal-correlation measurement results of the sampled points across the historical frames ($\mathbf{z}_t$ in Eq. 4).

The temporal-consistency measurement results of the sampled query units across all historical frames are shown in Fig. 11. The first two rows corresponds to the target units, while the remaining rows correspond to background units. It can be observed that, the temporal aggregation weights of different sampling points exhibit distinct changing trends. This also highlights the role of TCA in collecting and aligning relevant informative features.

After obtaining the temporal-consistency measurements for each unit on historical frame, we perform a weighted fusion according to these measurements (*i.e.* aggregation in TCA), thereby merging the $t-1$ historical feature tensors into a single feature tensor. This tensor reflects the final, complete spatiotemporal feature focusing result, as shown in Fig. 12. For comparison, the bottom row of Fig. 12 visualizes the original target-frame features. On the one hand, this comparison provides an intuitive view of the spatiotemporal focusing results for each query unit on the target frame; on the other hand, it indirectly reveals how the spatiotemporally focused historical frames help enhance the target features on the target frame.

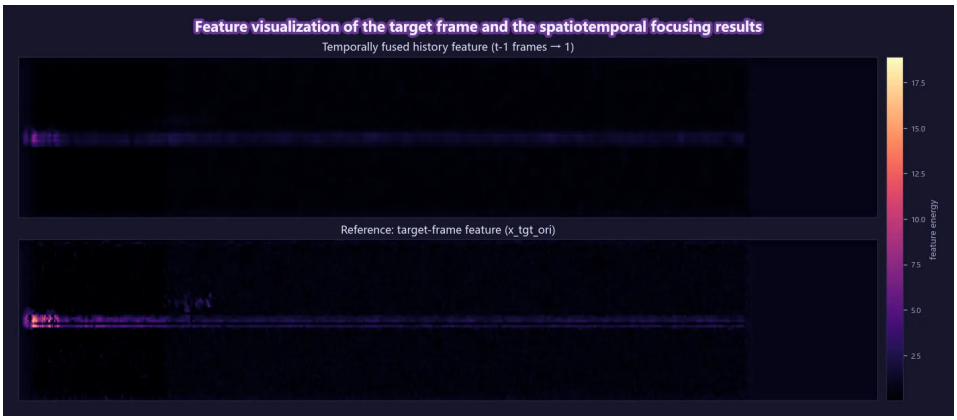


Figure 12: Feature visualization of the target frame and the spatiotemporal focusing results (the result of Eq. 6).

In TFoC, trajectory-consistency prediction is treated as an auxiliary classification task parallel to the main segmentation task. Fig. 13 presents a visual comparison between the predicted trajectory-consistency result (upper row) and the ground-truth trajectory mask (lower row), indicating that the trained TFoC model can effectively learn to predict target motion trajectories.

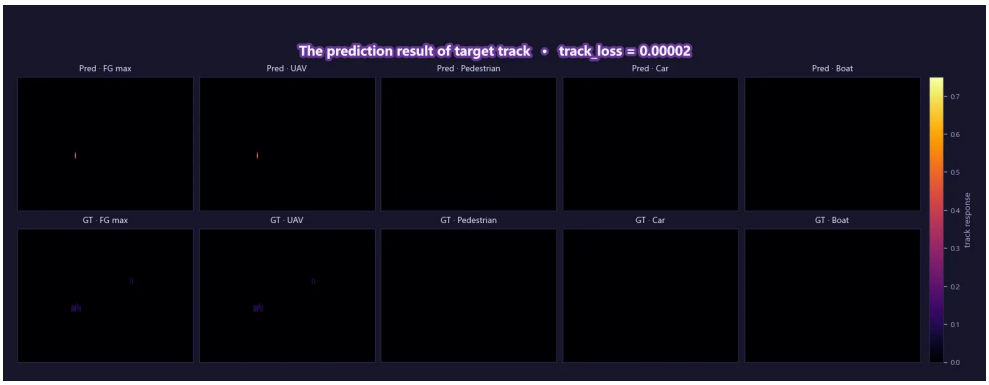


Figure 13: The prediction result of target track. Above: the predicted trajectory-consistency result. Below: the ground-truth trajectory mask. The track_loss value shown in the figure are the loss function values after TFoC training.

The final output of TFoC-Net is shown in Fig. 14. The first row presents the raw RD tensor, the second row shows the ground-truth mask of the target locations on the target frame, and the third row shows RSS predictions produced by TFoC-Net. Target regions are highlighted with cyan boxes, and both the ground-truth and predicted masks use different colors to indicate different target categories. As illustrated in Fig. 14, thanks to the spatiotemporal focusing mechanism, TFoC-Net accurately localizes targets and correctly recognizes its category.

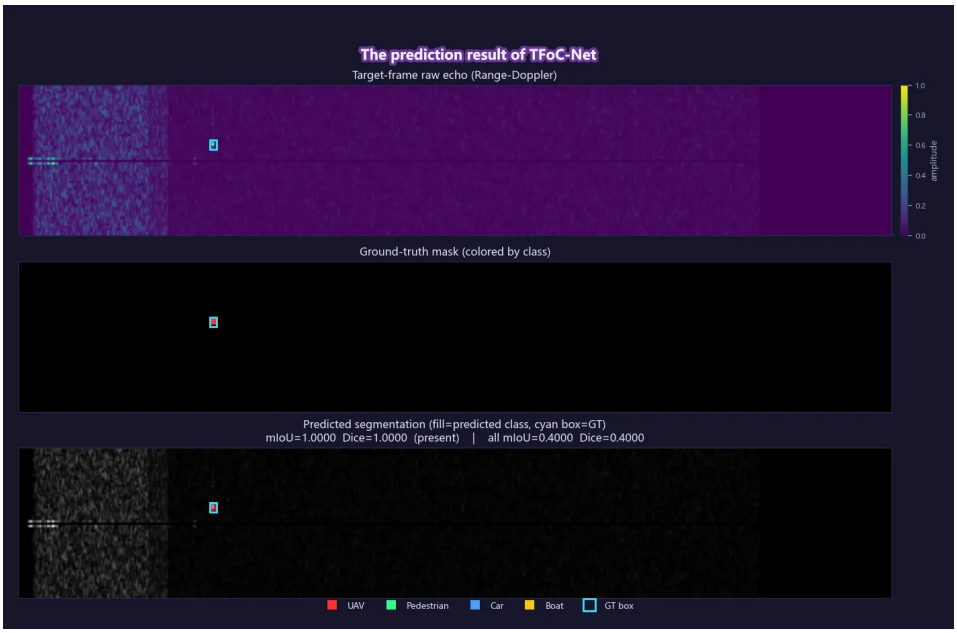


Figure 14: The prediction result of TFoC-Net.

More videos demonstrating the dynamic principles of TFoC have been packaged and uploaded as supplementary attachments.

F BLOCK-LEVEL SAF

In Section 3.1, we discuss the SAF implemented by anchoring individual units on the target frame. In this section, we extend single-unit anchoring to block regions to realize block-level SAF with an expanded receptive field range in historical frames. Fig. 15 illustrates block-level SAF.

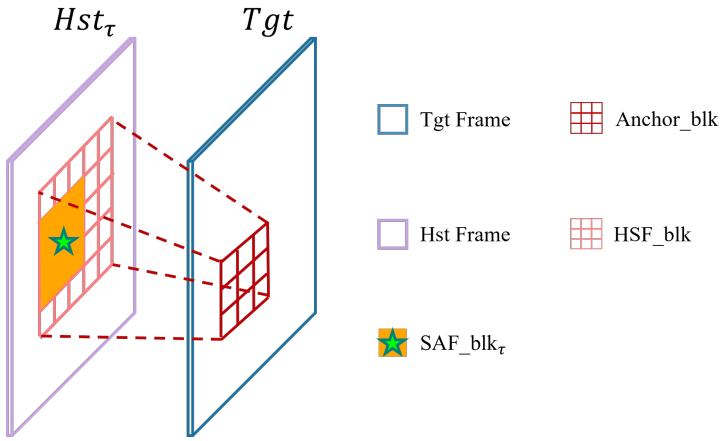


Figure 15: The illustration of block-level SAF.

It can be observed that after changing the search anchor from a single unit to the block level, TFoC performs semantic alignment calculations between the entire block data pattern and the patterns of each block in the historical frames. Since targets exhibit a certain degree of extensibility on the RD tensor, block-level SAF may better capture the overall pattern or distribution of targets. Compared to unit-level SAF, the alignment constraint is slightly relaxed. Accordingly, Eq. 10 is revised to:

$$\text{SAP}: \; P\left(j^\star = j^\dagger_{t,b} \mid \mathbf{x}^\star_b,\; \mathbf{x}^\dagger_{t,b}\right). \tag{24}$$

Where $j^\star$ denotes the class of each block within target frame, and $b = 1, 2, ..., B$ is the index of blocks (each block contains $k \times k$ units) on the historical frame mapped from each unit on the target block. HSF_blk is defied as:

$$\mathcal{H}_{t-blk} = \{\{\mathbf{x}^\dagger_{t,n,b}\}_{n=1}^{N}\}_{b=1}^{B} \tag{25}$$

From a similar derivation process, a variant of Eq. 2 for block-level SAF can be derived.

$$\text{SAF-blk: } \hat{\mathbf{H}}_{blk} \doteq \left\{\sum_{b=1}^{B} a_{t,b} \cdot \mathbf{x}^\dagger_{t,b}\right\}_{t=1}^{\tau}, \quad \text{where } a_{t,b} = \frac{\exp(A_{t,b}/\sigma_{\text{spa}})}{\sum_{m=1}^{B}\exp(A_{t,m}/\sigma_{\text{spa}})}. \tag{26}$$

In addition to utilizing block-level SAF, TFoC retains exactly the same procedures as the unit-level network. For differentiation from the previous TFoC network, we denote the TFoC network with block-level SAF as TFoC-blk.

F.1 Experiment of TFoC-blk

In the initial design of TFoC-blk, we believed that by substituting the original pairwise semantic consistency metric with a block-wise metric, it can properly fit the spectral broadening of radar data . With this design, the model can match small local data distributions (or patterns) on target frame to various patterns on historical frames and realize superior function of target feature gimbal. Is this assumption valid? The experiments in this section seek to verify it.

F.1.1 Training Setup for TFoC-blk

Hyperparameter: In the experiments presented below, the proposed TFoc-blk model adopts default hyperparameters: spatial-focusing "hardness" control factor σ_{spa} is set to 0.9, and the temporal decay factor σ_{temp} is set to 0.95. The search window size and input time-series length of the model vary across different datasets. On CARRADAR-RAC shown in Fig. 16, the TFoc-blk model takes 9 frames as input ($t = 9$). On the KuRALS dataset shown in Tab. 7, CW defaults $t = 7$ and $k = 5$, as well as PD defaults $t = 9$ with $k = 3$.

F.1.2 IMPACT OF SAF ON THE PERFORMANCE OF TFOC-BLK

As with TFoC, we investigate the performance effect of SAF on TFoC-blk in terms of search size and temperature factor, denoted as N in Eq. 25 and σ_{spa} in Eq. 26

Exploration of Search Size. Similar with TFoC, Search Size determines the scope with in which SAF performs semantic alignment matching, and directly affects the performance of TFoC_blk.TFoC-blk uses a $N = k \times k$ block on the target frame as anchor block, and its effective search domain is the aggregate coverage obtained by centering a block-sized window at every unit within that anchor block (e.g., a 3×3 anchor block induces an effective search region 5×5). Fig. 16 illustrates how the performance varies on the CARRADAR-RAC dataset with different anchor block sizes.

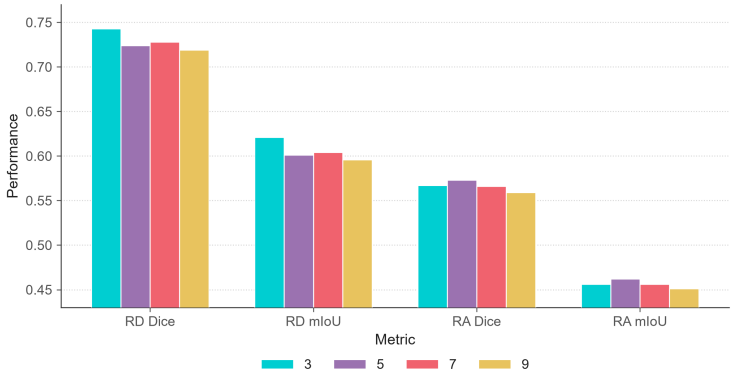


Figure 16: TFoC-blk performance trends under different search sizes on CARRADA-RAC.

It can be seen that TFoC-blk strengthens candidate competitiveness as k increases, and perform best on RD at k=3 and on RA at a moderately larger k=5. This is because RD requires tight Doppler alignment while RA needs a slightly wider search range to accommodate angular ambiguity. Nevertheless, enlarging the block search-size range fails to bring performance gains to TFoC-blk. Larger search-size receptive fields introduce more clutter and noise interference, resulting matching ambiguity and block-level spatial smoothing outweigh the benefit of a larger receptive field.

The Hardness of Spatial Focusing. Similar with TFoC, the hardness of SAF governs how rigorously target frame units filter semantically consistent information across different historical frames. σ_{spa} is the control factor for tuning the *hardness* of spatial semantic focusing (as difined in Eq. 26). As mentioned above, when $\sigma_{\text{spa}} \to 1$, the control strength weakens, allowing more block information from historical frames to be incorporated to realize spatial semantic focusing. When $\sigma_{\text{spa}} \to 0$, the control strength becomes stronger, and only a small number of historical frame blocks with high correlation are retained.

In `The Hardness of Spatial Focusing` of Sec. 4.1, we had an in-depth discussion of the hardness of spatial focusing for TFoC and most conclusions can be applied and transferred to TFoC_blk. In this way, we perform a preliminary exploration for TFoC_blk in this section. Tab. 6 presents the results.

Table 6: Performance of different hardness of spatial focusing.

Method	#Param.	RD-view (%)		RA-view (%)		Global
		mIoU	mDice	mIoU	mDice	Avg.
TFoC-blk w/ hard-selection	5.1M	56.7	68.4	43.4	53.3	55.5
TFoC-blk (default)	5.1M	**60.1**	**72.4**	**46.2**	**57.3**	**59.0**

`TFoC-blk w/ hard-selection` enforces the strictest selection of semantically-aligned blocks in historical frames at $\sigma_{\text{spa}} = 0$. As the TFoC-blk enforces spatiotemporal alignment between patches (with $k \times k$ units) in the search region of historical frames and anchor blocks, this extreme hard alignment impairs gradient back-propagation and thus performs worse than cases with moderately relaxed alignment constraints, *i.e.*, `TFoC-blk (default)`. This would serve as one of the factors to be considered in future work.

F.1.3 SOTA COMPARISON

The KuRALs dataset consists of the CW subset with poor temporal continuity and the PD subset with good temporal continuity, making it reasonably representative. Therefore, we adopt the KuRALs dataset to compare the SOTA performance of TFoC-blk.

Table 7: Performance comparison on KuRALS (KuRALS-CW: scanning mode; KuRALS-PD: tracking mode).

Method	#Param.	KuRALS-CW (%)			KuRALS-PD (%)		
		mIoU	mDice	Global avg.	mIoU	mDice	Global avg.
FCNLong et al. (2015)	134.3M	33.3	39.3	36.3	31.3	39.3	35.3
U-Net Ronneberger et al. (2015)	17.3M	47.2	56.5	51.9	46.6	58.7	52.7
DeepLabv3+Chen et al. (2017)	58.8M	50.5	58.8	54.7	53.3	66.3	59.8
HRNetSun et al. (2019)	65.8M	51.3	59.0	55.2	54.3	66.9	60.6
Swin-T Liu et al. (2021)	59.8M	51.9	59.1	55.5	52.5	63.7	58.1
SegFormer-B0 Xie et al. (2021)	3.7M	44.0	53.0	48.5	52.7	64.6	58.7
SegFormer-B1 Xie et al. (2021)	13.7M	48.8	59.2	54.0	52.6	65.0	58.8
RSSNetKaul et al. (2020)	10.1M	<u>56.8</u>	**67.5**	**62.2**	54.2	66.9	60.6
KuRALS-Net Li et al. (2026)	**1.2M/1.1M**	55.8	66.3	<u>61.1</u>	60.1	72.8	66.5
TMVA-Net^{sv}Ouaknine et al. (2021a)	**1.2M**	-	-	-	52.9	63.1	58.0
PKCIn-Net^{sv}Zhang et al. (2023b)	**1.2M**	-	-	-	56.7	65.9	61.3
TARSS-Net_S^{sv}Zhang et al. (2024)	**1.2M**	-	-	-	53.2	63.8	58.5
TARSS-Net_D^{sv}Zhang et al. (2024)	<u>1.3M</u>	-	-	-	58.4	67.1	62.8
TFoC-blk^{sv} (ours)	1.8M	55.9	63.1	59.5	<u>68.3</u>	**80.2**	<u>74.3</u>

As can be observed from the experimental results in Tab. 7, TFoC-blk fails to achieve SOTA performance on the CW dataset. However, on the PD dataset, it achieves semantic consistency alignment across temporal frames via block-wise SAF and obtains superior performance than existing RSS models.

F.1.4 EXPERIMENT SUMMARY OF TFOC-BLK

The experimental results are inconsistent with our expectations: TFoC-blk underperforms TFoC across all tests. In the `Exploration of search size`, TFoC consistently outperforms TFoC-blk, especially on RA. In the `SOTA Comparison`, TFoC-blk gains better performance on KuRALs-PD than on KuRALs-CW, and this experimental conclusions of TFoC-blk on KuRALs are consistent with those obtained from TFoC. However, TFoC-blk also achieves slightly inferior performance than TFoC.

Although by adopting block-wise semantic consistency measurement, TFoC-blk could effectively mitigates the risk of single anchor units aligning to isolated clutter or noise compared with TFoC, experiments on TFoC reveal that a larger searchsize introduces more clutter and noise, which leads to performance degradation. For the datasets in this paper, radar targets are small and highly sparse. In other words, a small search window can fully contain all target information. With the same value of k, TFoC-blk defines a larger range than TFoC at the stage of search domain definition. Therefore, it fails to leverage the advantage of block-wise target matching, instead, the enlarged search range introduces more irrelevant interference, leading to a loss of performance gain. In this situation, pixel-wise TFoC preserves finer spatial detail and suppresses weak correspondences more effectively than block-wise competition.

In summary, we maintain that TFoC-blk cannot achieve its full potential on current datasets, and its performance on datasets with more significant target spectral broadening remains promising.