SCIENCE CHINA
Information Sciences

• Supplementary File •

Towards low-altitude UAV detection in urban radar clutter: a benchmark and spatio-temporal channel transfer network

Jianhong ZHOU[1], Feng KE[1*], Yikui ZHAI[2*], Xueqiang ZHENG[3], Ziyi JIANG[1], Haolin LV[4] & Xiu Yin ZHANG[1]

[1]*School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China*
[2]*School of Electronic and Information Engineering, Wuyi University, Jiangmen 529020, Guangdong, China*
[3]*College of Communications Engineering, Army Engineering University of PLA, Nanjing Postcode, China*
[4]*Future Aviation Technology (Jiangmen) Co., Ltd., Jiangmen, Guangdong 210000, China*

Appendix A UAVRadar dataset

To bridge the gap in research on UAV radar target detection in urban environments and to enhance low-altitude urban security perception and protection capabilities, this paper presents UAVRadar, an RTK-annotated benchmark dataset for UAV radar target detection in urban environments. This section first introduces the data collection process in Subsection A.1, then describes the data annotation procedure in Subsection A.2, and finally presents detailed dataset statistics in Subsection A.3.

Appendix A.1 Data collection

Appendix A.1.1 *Collection workstation*

The data-acquisition workstation comprises two Texas Instruments (TI) evaluation boards and a Hikvision camera (see Figure A1(a)). Each board integrates an IWR6843-ISK millimeter-wave radar sensor with a DCA1000-EVM data-capture module. The IWR6843-ISK employs a remote antenna in a 3 transmit (TX)/4 receive (RX) MIMO configuration, operating over 60-64 GHz with a 120° azimuth field of view (FoV) and a 30° elevation FoV. The DCA1000-EVM acquires raw analog-to-digital converter (ADC) samples in real time and streams them to storage. To capture both azimuthal and elevational information, the two boards were deployed in horizontal and vertical orientations to mitigate the remote antenna's limited elevation coverage. The camera provides synchronized imagery to facilitate subsequent visualization and analysis. The target was a DJI Matrice 300 UAV equipped with a real-time kinematic (RTK) module (see Figure A1(b)). Unless otherwise stated, the experiments reported here use only the horizontally deployed radar; camera imagery and data from the vertically deployed radar are omitted and reserved for future study.

Appendix A.1.2 *Radar configuration*

In our implementation, we employ a time-division multiplexing MIMO acquisition scheme with a 3-TX/4-RX configuration. The chirp duration is $T_c = 380\ us$, and each frame comprises $N_c = 255$ chirps. The FMCW slope is $S = 9.994$ MHz/us, and the ADC sampling rate is $f_s = 6250$ KSps. With $N_s = 512$ samples per chirp, the sweep bandwidth is $B = N_s S/f_s \approx$ 818.71 MHz, which corresponds to a range resolution of $R_{res} = c_0/(2B) \approx$ 0.1832 m and a maximum unambiguous range of $R_{max} = f_s c_0/(2S) \approx$ 93.8 m, where c_0 denotes the speed of light. Range and radial velocity are obtained via range fast Fourier transform (FFT) and Doppler-FFT processing, respectively, following the standard FMCW signal model. For angle estimation, we combine TX_1 and TX_3 with all receive antennas to synthesize an 8-channel virtual uniform linear array and apply an angle-FFT beamformer, achieving an angular resolution of approximately 14.3°. We exclude TX_2 from angle processing because its slight elevation offset relative to the other transmit elements induces out-of-plane array distortion (see Figure A1(a)); without calibration, this mismatch degrades angular estimation performance. Additional radar parameters are summarized in Table A1.

Appendix A.2 Data annotation

To obtain accurate labels for model training and evaluation, we annotate the data using RTK measurements. Specifically, we compute the target range and azimuth relative to the radar and localize the target in the range-azimuth (RA) coordinate system via a nearest-neighbor search. The resulting annotations are then converted into a representation compatible with the model input. Following the work in [1], we construct a confidence map (ConfMap) in the RA plane to encode the target location. In the ConfMap, the values in the vicinity of the target are modeled as a Gaussian distribution with its mean centered at the target position and its variance determined by the target category and scale. Prior to ConfMap generation, the RTK measurements are transformed into the RA coordinate system.

* Corresponding author (email: fengke@scut.edu.cn, yikuizhai@163.com)

Table A1 Configuration and parameters of TI's IWR6843-ISK FMCW radar

Initial frequency (f_c)	60.25 GHz	Sweep bandwidth (B)	818.71 MHz
Sweep slope (S)	9.994 MHz	Wavelength (λ)	0.00498 m
Sampling frequency (f_s)	6250 KSps	Range resolution (R_{res})	0.1832 m
Number of chirps in one frame (N_c)	255	Max operating range (R_{max})	93.8 m
Number of samples of one chirp (N_s)	512	Velocity resolution (V_{res})	0.0257 m/s
Number of transmitters, receivers (N_{TX}, N_{RX})	3, 4	Max operating velocity (V_{max})	3.2763 m/s
Sampling period of one chirp (T_c)	380 us	Angle resolution (θ_{res})	$\sim$14.3°
Frame rate (f_F)	10 FPS	Max operating angle (θ_{max})	$\sim$120°

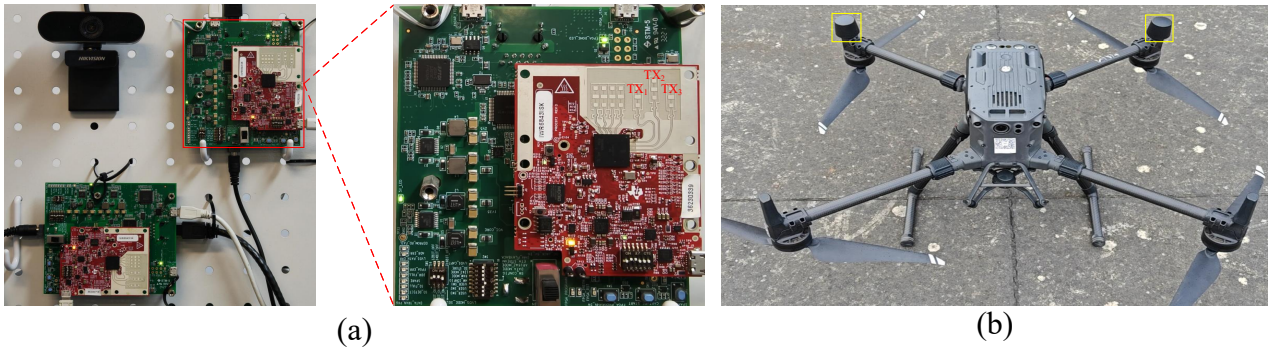

Figure A1 Hardware configuration and structural schematic. (a) The left panel shows the collection workstation, comprising a Hikvision camera and two TI evaluation boards. The upper board is mounted horizontally, whereas the lower board is mounted vertically. The right panel presents a magnified view of the region outlined by the red solid box, which clarifies the TI evaluation-board assembly: the red board corresponds to the IWR6843-ISK millimeter-wave radar sensor, and the green board corresponds to the DCA1000-EVM capture board. (b) The DJI Matrice 300 UAV is shown; the area enclosed by the yellow solid boxes indicates the integrated RTK module.

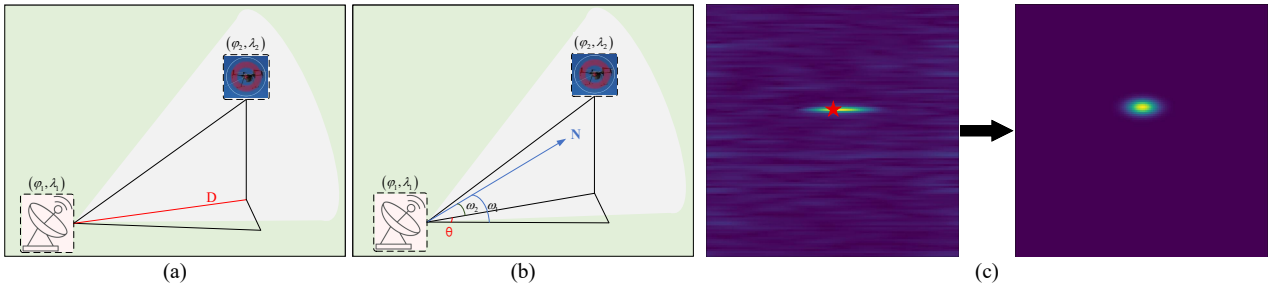

Figure A2 UAV-radar geometry computation and ConfMap label generation. (a) Range estimation. (b) Azimuth estimation relative to the radar heading. (c) Annotation generation from the estimated range and azimuth.

This mapping comprises two steps: (1) estimating the radar-UAV range (see Figure A2(a)) and (2) computing the UAV azimuth (see Figure A2(b)) θ with respect to the radar heading. The UAV-radar distance is estimated using the Haversine formula:

$$D = 2R \times \arcsin\sqrt{\sin^2\frac{\Delta\varphi}{2} + \cos\varphi_1 \times \cos\varphi_2 \times \sin^2(\frac{\Delta\lambda}{2})}, \tag{A1}$$

where R denotes the Earth's radius, assumed to be 6,371,009 m. The radar and UAV coordinates are (φ_1, λ_1) and (φ_2, λ_2), respectively. Here, $\Delta\varphi$ and $\Delta\lambda$ represent the differences in latitude and longitude, defined as $\Delta\varphi = \varphi_2 - \varphi_1$ and $\Delta\lambda = \lambda_2 - \lambda_1$. To compute the azimuth, we adopt the great-circle bearing formula to obtain the initial bearing along the great-circle path from the radar to the UAV, expressed as:

$$\omega_2 = \arctan\frac{\sin\Delta\lambda}{\cos\varphi_1 \times \tan\varphi_2 - \sin\varphi_1 \times \cos\Delta\lambda}. \tag{A2}$$

Because the UAV azimuth relative to the radar heading cannot be measured directly, we record the angle ω_1 between the radar heading and true north before data acquisition. The relative azimuth is then computed as $\theta = \omega_2 - \omega_1$. Finally, these estimates are used to label each radar frame and to construct the corresponding ConfMap for model training (see Figure A2(c)).

Appendix A.3 Data statistics

After data cleaning, filtering, and annotation, 79 sequences were retained, comprising a total of 2,370 seconds of UAV tracking signal data. The UAVRadar dataset was collected at multiple locations and across different time periods, including representative urban scenarios such as Jiangmen High-Tech Zone, South China University of Technology, and Wuyi University. Data acquisition covered

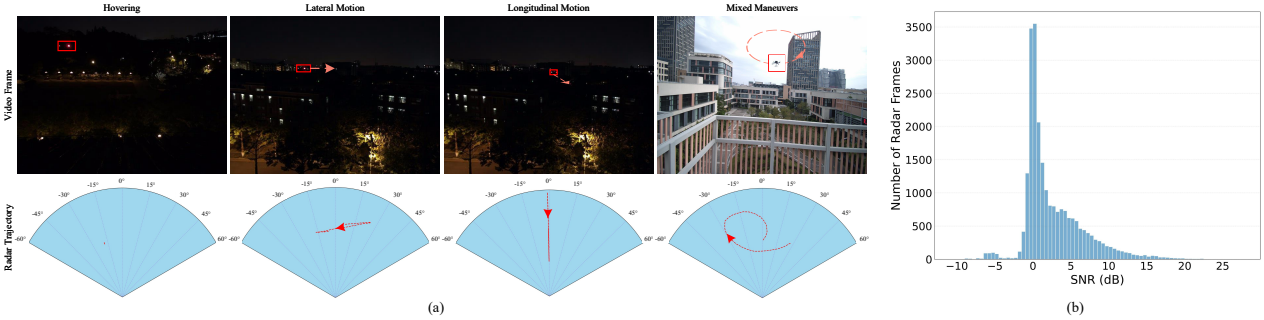

Figure A3 Overview of the UAVRadar dataset. (a) Example samples from the UAVRadar dataset and their corresponding target motion trajectories. Four representative motion patterns are illustrated: hovering, lateral motion, longitudinal motion, and mixed maneuvers. The top row presents visible-light images of the data acquisition scenes, where red arrows indicate the directions of target motion. The bottom row shows UAV motion trajectories reconstructed from RTK data, where red dots represent the target trajectory over time. (b) Distribution of target SNR in the UAVRadar dataset.

Table A2 Data split of the UAVRadar dataset.

Split	Longitudinal	Lateral	Hovering	Mixed	Total	Proportion
Training	4,200	5,100	3,900	6,000	19,200	81.0%
Testing	1,200	1,800	1,500	–	4,500	19.0%
Total	5,400	6,900	5,400	6,000	23,700	100%

Table A3 Comparison of radar UAV detection datasets.

Dataset	Scenarios	Band	Data Types	Task	Sequences	Data Size	Categories	Annotation	Max Range
CubeDN	Day/Night	mmWave	R-V-A-E	3D Localization	41	64,000	1	RTK	15m
mDrone	Day	mmWave	R-V	3D Localization	100	100 min	1	Vicon	4m
LSS-FMCWR	Day	Ku+L	mDS	Classification	114	-	6	Manual	12m
LFMCWR	Day	K+L	mDS	Classification	-	10,567	6	Manual	10m
LSS-UAV	Day/Night	S	R-V/mDS	Detection/Classification	13	30 min	2	GPS	654m
RDRD	Day/Night	X	R-V/mDS	Detection/Classification	-	17,485	3	Manual	3596m
MMMAUD	Day	mmWave	Multimodal	Detection/Tracking/Classification	50	28 min	5	Leica	20m
UAVRadar (ours)	Day/Night	mmWave	R-V-A	Detection/Tracking/Classification	79	23,700	3	RTK	93.8m

multiple times of day, including early morning, noon, and evening, and captured radar echoes under various weather conditions, such as clear skies, strong winds, and light rain, as well as at different flight speeds. These diverse factors enable UAVRadar to comprehensively characterize the radar echo properties of UAVs under different operating conditions in urban environments, thereby providing a reliable foundation for training and evaluating UAV radar target detection models. To illustrate the diversity of UAV motion patterns in the dataset, Figure A3(a) presents representative trajectories corresponding to different motion types. These visualizations provide a more intuitive understanding of UAV motion characteristics and echo variations under different flight modes. In addition, Figure A3(b) shows the distribution of target SNR. The results indicate that, in the collected scenarios, SNR values are mainly concentrated around 0 dB, suggesting that the dataset contains a large number of weak-echo targets and further highlighting the challenges of UAV radar target detection. In terms of data partitioning, the UAVRadar dataset is divided into four flight modes: longitudinal flight, lateral flight, hovering, and mixed maneuvers. Mixed maneuvers refer to flight trajectories that do not strictly belong to the first three categories and typically involve composite or more complex motion patterns. To evaluate model generalization to unseen acquisition scenes, we adopt a strict scene-disjoint partitioning protocol. Specifically, the training and test sets were collected at distinct locations and scenes, with no acquisition scene shared between the two subsets. Thus, all results reported in this study are obtained under a cross-scene evaluation setting within the UAVRadar benchmark. Detailed data partitioning results are presented in Table A2. All training and test data, together with their annotation information, will be publicly released.

Appendix A.4 Comparison with existing Radar UAV detection datasets

Existing radar datasets for UAV sensing support localization, detection, tracking, and classification. Early datasets primarily focused on high-precision localization in controlled environments. For example, CubeDN [2] and mDrone [3] use millimeter-wave (mmWave) radar with external motion-capture or RTK systems for precise three-dimensional UAV localization. However, both datasets are largely limited to localization and have relatively short sensing ranges. For low-altitude, slow-moving, and small UAVs, the LSS datasets [4–6] provide multiband radar measurements from diverse environments to support longer-range target detection and UAV type identification. RDRD [7] further broadens task coverage by supporting multiclass classification using range-velocity (RV) and mDS representations but lacks continuous trajectory-level annotations. More recent multimodal datasets, such as MMAUD [8], integrate radar, visual, and LiDAR data to provide richer semantic information and support joint detection, tracking, and classification. However, they rely on auxiliary sensors, and their sensing ranges remain relatively limited. In contrast, the proposed UAVRadar dataset is designed for

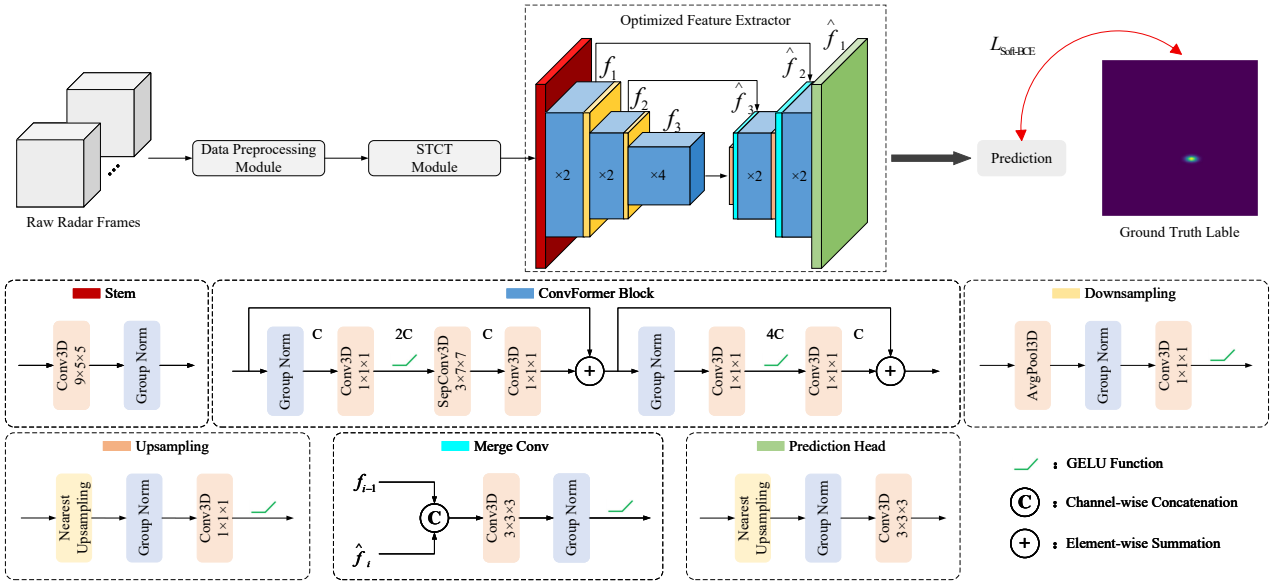

Figure B1 The overall architecture of STCT-Net comprises a data preprocessing module, an STCT module, and an optimized feature extractor. The raw radar frame sequence is first processed by the data preprocessing and STCT modules to extract micro-temporal features and construct an input representation suitable for subsequent network learning. All components shown in the lower portion of the figure, including the stem, ConvFormer block, downsampling and upsampling modules, Merge Conv, and prediction head, constitute the optimized feature extractor. The resulting features are then fed into the optimized feature extractor, which adopts a U-shaped encoder-decoder architecture. In the encoder, the input features are progressively downsampled to generate multi-scale feature maps, denoted as f_1, f_2, and f_3. In the decoder, these feature maps are progressively upsampled and fused, through skip connections, with the high-resolution features from the corresponding encoder stages, yielding refined feature maps $\hat{f}_1$, $\hat{f}_2$, and $\hat{f}_3$. Finally, the decoder output is passed through a prediction head to produce the final detection results, and the proposed Soft-BCE loss function is employed for supervised training using the ground-truth labels.

low-altitude UAV sensing in complex daytime and nighttime scenarios. It provides temporally continuous RTK supervision and adopts a unified R-V-A representation to support integrated sensing tasks across heterogeneous radar configurations and at longer ranges. Table A3 provides a detailed comparison of these datasets.

Appendix B Method

For UAV radar target detection, this paper proposes a UAV detection framework termed STCT-Net. The framework comprises three components: a data preprocessing module, an STCT module, and an optimized feature extractor. Among these, the data preprocessing module and the STCT module are specifically designed for input feature enhancement and spatio-temporal information modeling, respectively. Specifically, during the data preprocessing stage, a bin-based PCA dimensionality reduction strategy is introduced along the chirp dimension of the RA spectrum to replace the conventional uniform sampling approach. Compared with conventional methods, this strategy more effectively highlights the critical dynamic information of weak and small targets, thereby enhancing feature saliency and providing a more compact, information-dense input representation for subsequent network learning. In the STCT module, a chirp merging mechanism is further introduced. This mechanism explicitly decouples the modeling of spatial and temporal features, enabling the network to preserve key temporal characteristics in target echoes more effectively while extracting spatial information, thereby improving its ability to characterize the dynamic properties of UAV targets. For feature extraction, this study adopts the ConvFormer paradigm, based on the MetaFormer [9] framework. This paradigm offers strong feature extraction capability and stable training performance. To accommodate the characteristics of radar target detection, it is extended from a 2D architecture to a 3D one, and its key components are selectively replaced and optimized to improve its suitability for radar data representation. In addition, given the weak-target characteristics of UAVs in radar sensing, a task-adapted training loss function, Soft-BCE, is proposed. During training, this loss function places greater emphasis on modeling the responses of target regions while enhancing the suppression of background noise, thereby improving weak-target detection performance. Owing to the coordinated design of these modules, the proposed STCT-Net achieves excellent performance in UAV detection task. Figure B1 presents the overall architecture of the proposed method. The following sections describe the design motivation and implementation details of each module in accordance with the raw-data processing pipeline.

Appendix B.1 Data preprocessing module

In order to enhance weak targets for UAV radar target detection, a data preprocessing method is proposed, as illustrated in Figure B2a. Within a single coherent processing interval, the FMCW radar transmits N_{C} chirp signals to form one radar frame. After reception, mixing, and ADC digitization, the echo signals are organized into a raw data cube of size $(N_{\mathrm{C}}, N_{\mathrm{S}}, N_{\mathrm{RX}})$, where N_{S} denotes the number of ADC sampling points per chirp and N_{RX} denotes the number of virtual receive channels. During preprocessing, the direct current (DC) component is first removed to suppress static clutter. The data cube is then subjected to a two-stage FFT. Specifically, the first stage is performed along the fast-time dimension to extract range information, yielding R range bins, while the second stage is performed along the virtual antenna dimension to resolve azimuth information, yielding A angle bins. After these operations, the

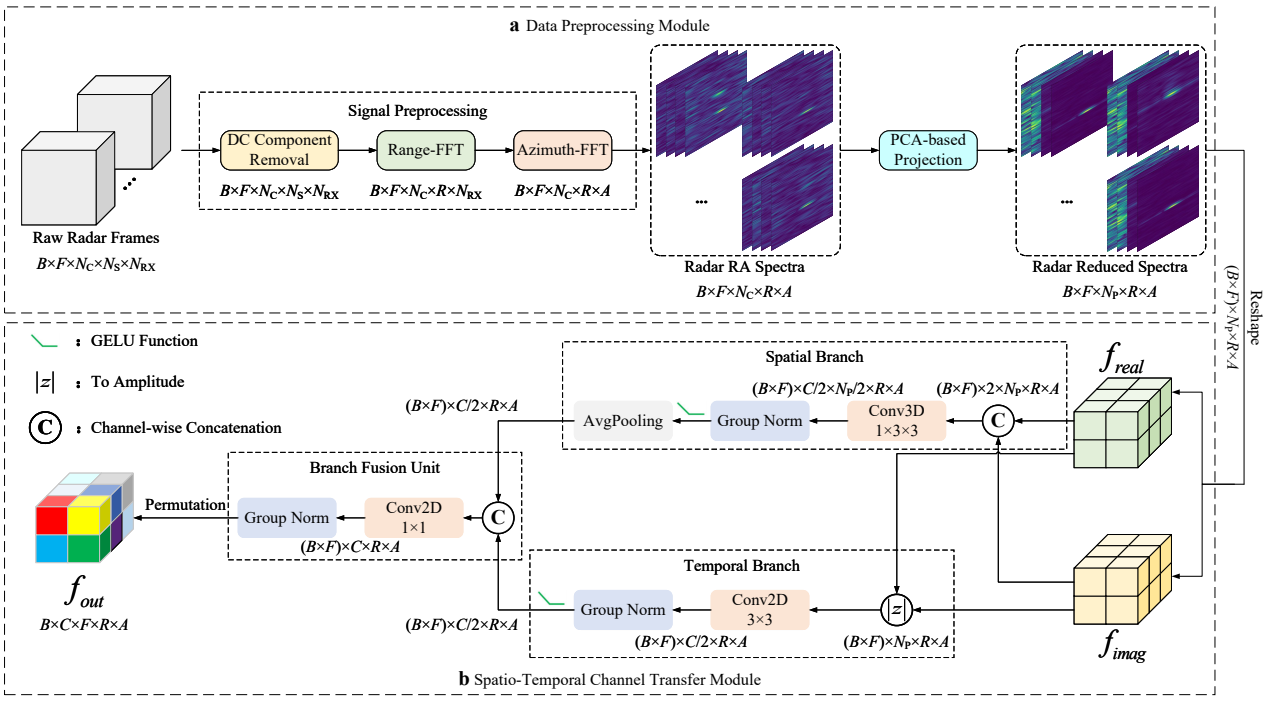

Figure B2 Conceptual illustration of the proposed modules. (**a**) The proposed data preprocessing module is designed to preserve the dynamic information contained in the chirp sequence to the greatest extent possible. First, DC component removal is applied to the raw radar frames to effectively suppress static clutter. Spatial information is then extracted through a two-stage FFT procedure: in the first stage, range-FFT is performed along the fast-time dimension N_{S} to map the signal to the range dimension R; in the second stage, azimuth-FFT is performed along the virtual antenna dimension N_{RX} to further map the signal to the angle dimension A. After these operations, the raw data are reconstructed into a series of radar RA spectra. To reduce data dimensionality while retaining critical dynamic features, a PCA-based projection is further applied along the chirp dimension, compressing the original N_{C} chirps into N_{P} principal components. Ultimately, a dimension-reduced radar spectrum tensor of size $B \times F \times N_{\mathrm{P}} \times R \times A$ is obtained and used as the input to the STCT module. (**b**) The proposed STCT module is designed for chirp merging and consists of three core components: a temporal branch, a spatial branch, and a branch fusion unit. The module achieves effective feature fusion through a decoupled spatio-temporal feature extraction mechanism. Specifically, the temporal branch models the temporal dependencies of each bin to capture the dynamic characteristics of the target, whereas the spatial branch focuses on extracting spatial texture information from the RA spectrum. The outputs of the two branches are then integrated through the branch fusion unit, thereby generating a unified and more discriminative feature representation for each frame.

raw data are reorganized into N_{C} independent RA spectrum slices. Unlike conventional uniform sampling strategies or standard Doppler transformation methods, the proposed method focuses on targeted processing of the chirp dimension, N_{C}. Specifically, PCA is independently performed on each RA bin for dimensionality reduction and feature enhancement. This strategy effectively compresses and aggregates the micro-temporal feature energy contained in the chirp sequence into N_{P} principal components, thereby reducing data dimensionality while preserving the dynamic features that are most discriminative for target recognition. Under this scheme, bin-wise PCA not only significantly reduces the complexity of the input data but also better preserves the key temporal information associated with weak targets. Ultimately, this module outputs a PCA-reduced RA spectrum tensor of size (N_{P}, R, A), which serves as the input to the subsequent network.

Notably, a Hamming window is applied before the fast-time dimension FFT to minimize spectral leakage and suppress spurious peak interference, thereby enhancing range resolution. The Hamming window is defined as:

$$w_h(n) = 0.54 - 0.46 \times \cos\left(\frac{2\pi n}{N_h - 1}\right), \quad 0 \leqslant n \leqslant N_h - 1 \tag{B1}$$

where N_h denotes the length of the Hamming window, and n indicates the sample index. Before the virtual antenna dimension FFT operations, a Taylor window is used to optimize the dynamic range by controlling sidelobes, improving angular resolution. The Taylor window is expressed as:

$$w_t(n) = 1 + 2\sum_{m=1}^{M-1} F_m \cos\left[\frac{2\pi m}{N_t}\left(n - \frac{N_t}{2}\right)\right], \quad 0 \leqslant n \leqslant N_t - 1 \tag{B2}$$

where N_t denotes the length of the Taylor window, M represents the number of sidelobe orders, and F_m is the amplitude factor determined by the Taylor weighting coefficients, typically calculated based on the desired sidelobe level and M.

Appendix B.2 Spatio-temporal channel transfer module

In order to fully preserve and effectively exploit the temporal dynamics embedded in chirp features, we propose a novel chirp merging strategy. The strategy consists of two independent branches: a temporal branch and a spatial branch. The temporal branch is designed to model the temporal dynamics of each bin, thereby capturing temporal cues more comprehensively and characterizing the target's

micro-motion features. Specifically, the input data are first converted into magnitude representations, and the chirp dimension is then reshaped into the channel dimension to jointly encode the local intensity distribution in the RA spectrum and the temporal variations across chirps, thereby preserving dynamic information to the greatest extent possible. In contrast, the spatial branch focuses on extracting texture features from the RA spectrum by applying average pooling along the chirp dimension. Compared with max pooling, which is more sensitive to outliers and may result in information loss, average pooling more effectively suppresses noise and captures the overall trend of variation, making it better suited to the weak-target UAV radar detection scenario considered in this study. Finally, the features extracted by the two branches are integrated through a branch fusion unit.

Figure B2b illustrates the overall framework of the proposed spatio-temporal feature decoupling and branch fusion unit. The raw input is a complex-valued RA spectrum sequence with dimensions of $(B \times F) \times N_{\mathrm{P}} \times R \times A$, where B denotes the batch size, F denotes the number of frames, N_{P} denotes the number of retained principal components, and R and A represent the range and angle dimensions, respectively. Because the module processes the input independently on a frame-by-frame basis, it first extracts features from each frame, $f_i \in \mathbb{R}^{B\times N_{\mathrm{P}}\times R\times A}$, and then generates the corresponding fused representation. During single-frame feature extraction, the complex-valued RA spectrum is first decomposed into its real and imaginary components, denoted by f_{real} and f_{imag}, respectively. These two components are then concatenated along the channel dimension to form an input tensor of size $B \times 2 \times N_{\mathrm{P}} \times R \times A$, which is subsequently fed into the spatial branch for spatial texture extraction. Specifically, the spatial branch employs a 3D convolution with a kernel size of $1 \times 3 \times 3$ to model local spatial patterns, followed by group normalization (GN) and the Gaussian error linear unit (GELU) for feature normalization and nonlinear transformation. Average pooling is then applied along the chirp dimension to enhance the representation of global variation trends and to suppress noise introduced by outliers. In parallel, the temporal branch takes the complex magnitude $\left|f_{\mathrm{real}} + j \cdot f_{\mathrm{imag}}\right|$ as input to preserve temporal intensity variations across chirps. Specifically, a 3×3 2D convolution is first used to model the local temporal distribution within each chirp frame, followed by GN and GELU for feature normalization and nonlinear transformation, thereby generating a feature representation with the same dimensionality as the output of the spatial branch. The outputs of the two branches are then concatenated along the channel dimension and fused by a 1×1 2D convolution, followed by GN and GELU to further enhance the representational capability of the features. The resulting tensor constitutes the fused feature f_{out}. This process is repeated independently for each frame in the sequence, thereby generating high-quality frame-wise feature representations over the entire temporal range. Notably, because the statistical computation of GN is independent of the batch size, the batch dimension B and the frame dimension F can be merged, enabling more efficient parallel computation within a single forward pass. Based on empirical observations, the number of principal components retained after dimensionality reduction, N_{P}, is set to 32 in this study. It should be emphasized that the temporal branch is specifically designed to characterize fine-grained intra-frame dynamics across chirps, rather than to function as an explicit operator for modeling long-range temporal dependencies across radar frames. By establishing a deterministic correspondence between chirp positions and feature channels, the temporal-to-channel transformation preserves the structured organization of chirp-wise variations and facilitates the extraction of subtle motion-induced signatures within each frame. Importantly, this transformation preserves the inter-frame temporal axis, enabling the subsequent feature encoder to further capture higher-level spatio-temporal evolution across consecutive radar frames.

Appendix B.3 Optimized feature extractor

As illustrated in Figure B1, STCT-Net employs a feature extractor based on the ConvFormer framework within the MetaFormer paradigm and organized as a U-shaped encoder-decoder architecture. The input 3D radar features are first processed by a stem module to extract low-level representations. Successive stages of ConvFormer blocks and downsampling modules then progressively reduce the feature resolution while capturing high-level semantic information. The decoder takes the deepest encoder representation as input and progressively restores its spatio-temporal resolution through a series of upsampling operations. At each decoding stage, the upsampled features are fused with the corresponding encoder features through skip connections and subsequently processed by a Merge Conv module. Finally, the prediction head maps the fused representation to bin-wise classification and position-regression outputs, thereby establishing a complete processing pipeline consisting of encoding, decoding, multiscale feature fusion, and prediction.

To accommodate the 3D structure of radar data, the 2D convolution in the token mixer of each ConvFormer block is replaced with a 3D depthwise separable convolution. This modification enables the joint modeling of dependencies across the temporal, range, and azimuthal dimensions while reducing computational cost. Because the input has a relatively limited temporal extent, an asymmetric $3\times7\times7$ depthwise convolutional kernel is employed, with a kernel size of 3 along the temporal dimension and 7×7 in the RA plane. This configuration enlarges the receptive field in the RA plane while maintaining computational efficiency, thereby enhancing the extraction of structural features, including target contours, edges, and spatial distributions. At each encoder stage, ConvFormer blocks first extract spatio-temporal features, after which a downsampling module reduces the feature-map resolution to enlarge the effective receptive field. Following the design in Reference [10], GN and a $1 \times 1 \times 1$ 3D convolution are sequentially applied after each downsampling module to stabilize feature distributions across scales and adjust the channel dimensionality. The decoder progressively upsamples the features in reverse order and applies the same GN and $1 \times 1 \times 1$ 3D convolution for feature normalization and channel adjustment. The upsampled features are then fused with the corresponding encoder features at the same scale using a Merge Conv module. This module consists of a $3 \times 3 \times 3$ 3D convolution, GN, and a GELU activation. It integrates fine-grained spatio-temporal information from the encoder with high-level semantic information from the decoder, thereby mitigating the loss of spatial and temporal details caused by downsampling.

To improve the model's compatibility with 3D radar data and small-batch training, the normalization, activation, and channel-projection components of the ConvFormer block are modified accordingly. Processing 3D inputs substantially increases GPU memory consumption and computational cost, thereby constraining the feasible batch size. Layer normalization (LN) is therefore replaced throughout the network with GN, which is independent of batch-level statistics. In addition, StarReLU is replaced with GELU, which offers a smoother activation profile and non-monotonic behavior. A $1 \times 1 \times 1$ 3D convolution is used for channel projection, preserving the fully convolutional architecture and enabling the network to process inputs of varying sizes. At the input stage, inspired by the large-kernel Stem design in Reference [1], a $9 \times 5 \times 5$ 3D convolution aggregates dynamic information across consecutive frames while extracting local structures in the RA plane, thereby enlarging the initial receptive field. At the output stage, the Prediction Head,

comprising GN followed by a $3\times3\times3$ 3D convolution, jointly predicts the target class and positional parameters at each spatio-temporal bin.

Appendix B.4 Soft binary cross-entropy loss

Given that the supervision signal in this task is a Gaussian/elliptical heatmap constructed on the RA spectrum, as shown in Figure B3, the network prediction can be interpreted as estimating the target confidence for each bin, i.e., a soft Bernoulli-style confidence encoding. Specifically, $y = 1$ corresponds to the target center, $0 < y < 1$ denotes a smoothed transition region in the vicinity of the target center, and $y = 0$ represents the background. Under this label-encoding scheme, the objective of training is not to strictly fit the continuous response distribution over the entire Gaussian ellipse, but rather to accurately activate the target center in a complex noisy background while maintaining appropriate tolerance to deviations in the central neighborhood caused by heatmap discretization, target extent variations, and annotation uncertainty. Motivated by these considerations, we propose a Soft-BCE loss based on the standard BCE loss. Unlike standard BCE, which applies equally strong negative-sample penalties to all non-central locations, the proposed Soft-BCE treats only locations with $y = 1$ as positive samples, whereas all remaining locations with $y \in [0, 1)$ are regarded as negative samples with attenuated penalties. Specifically, a label-dependent weighting term, $(1 - y)^\theta$, is introduced to modulate the negative-sample loss, such that locations closer to the target center receive weaker negative penalties. In contrast, for locations farther from the target center and closer to the background, the penalty gradually approaches that of standard BCE. This design ensures strong supervision at the target center while effectively reducing excessive penalties on predictions that are close to the correct location but do not fall exactly at the center, thereby improving the model's robustness to slight positional offsets, target extent variations, and heatmap annotation errors. This idea is analogous to strategies used in heatmap-based keypoint detection, where Gaussian labels are employed to reduce the penalty in the neighborhood of the target center [11]. Accordingly, for an arbitrary bin, the Soft-BCE loss is defined as follows:

$$\mathcal{L}_{\text{Soft}} = \begin{cases} -\log(p), & \text{if } y = 1, \\ -(1-y)^\theta \log(1-p), & \text{if } y \in [0, 1), \end{cases} \tag{B3}$$

where $p \in (0, 1)$ denotes the target confidence predicted by the model at the current bin, $y \in [0, 1]$ denotes the corresponding heatmap label, and θ is a hyperparameter that controls the attenuation of the penalty assigned to negative samples. Across the entire RA spectrum, the training loss is defined as follows:

$$\mathcal{L}_{\text{Soft-BCE}} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \mathcal{L}_{\text{Soft}}(p_u, y_u), \tag{B4}$$

where Ω denotes the set of all bins in the RA plane. From the gradient expression, it follows that, when $y \in [0, 1)$, we obtain:

$$\frac{\partial L_{\text{Soft}}}{\partial s} = (1-y)^\theta p, \tag{B5}$$

where s denotes the corresponding logit. The term $(1 - y)^\theta$ directly modulates the gradient magnitude at non-central locations. Specifically, when a location is closer to the target center, its corresponding y value is larger, yielding a smaller $(1 - y)^\theta$ and thus substantially suppressing the negative gradient. In contrast, when a location is farther from the target center and approaches the background region, its y value decreases, and the corresponding gradient gradually approaches that of the standard BCE. For the center point, however, the loss remains $-\log(p)$, so the positive supervision at the center is preserved.

When $\theta = 0$, the loss reduces to the standard BCE, which applies conventional binary supervision to both center and non-center points. When $0 < \theta < 1$, the penalty decays more gradually, so non-central locations within the neighborhood of the center still retain a certain degree of supervision, providing a robust trade-off between boundary tolerance and center supervision. When $\theta = 1$, the negative-sample penalty decreases linearly with y. When $\theta > 1$, the penalty for locations near the center decreases more rapidly, indicating that the model is more tolerant of positional deviations within the target neighborhood. However, if θ is too large, supervision in the transition region may become insufficient, causing the predicted boundary to become overly loose. Based on empirical observations, we set $\theta = 1$ in this paper. Although this mechanism and label smoothing can both alleviate overly rigid supervision, they are fundamentally different: label smoothing directly modifies the target distribution, whereas Soft-BCE only reweights the negative-sample penalty in non-central regions while leaving the supervision at the center unchanged.

Appendix C Experiments

In this section, we conduct comprehensive experiments on the UAVRadar dataset to evaluate the effectiveness of the proposed STCT-Net framework. Specifically, Subsection C.1 describes the experimental settings, Subsection C.2 compares the proposed method with several state-of-the-art (SOTA) radar target detection models, Subsection C.3 presents the ablation study results, and Subsection C.4 provides further analysis and discussion of the detection performance.

Appendix C.1 Experimental settings

We respectively introduce the comparative methods and the detailed implementations in this subsection.

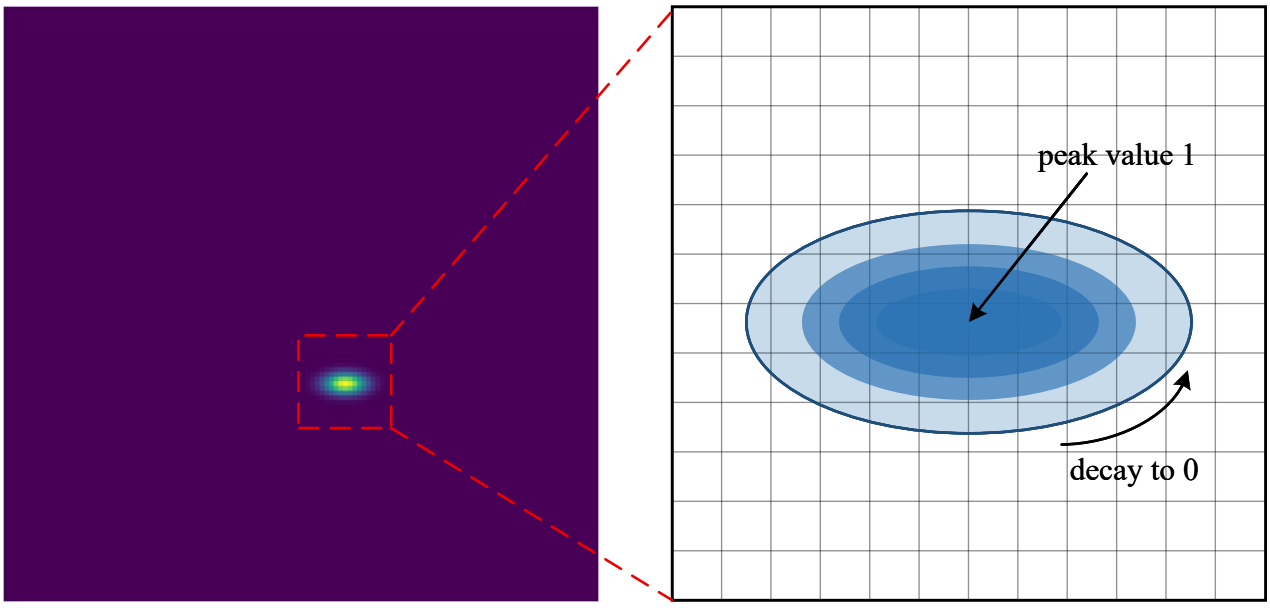

Figure B3 Schematic illustration of Gaussian heatmap labels. The target center is assigned a label value of $y = 1$, the surrounding transition region is encoded with soft labels satisfying $0 < y < 1$, and the background region is assigned a label value of $y = 0$.

Appendix C.1.1 *Comparative methods and metrics*

We compared several representative radar target detection methods encompassing diverse input paradigms and network architectures. RODNet-CDC [1], RODNet-HG [1], RODNet-HGwI [1], and E-RODNet [12] use multi-frame, multi-chirp radar inputs and convolution-based architectures. RAMP-CNN [13] combines multi-frame radar inputs with multi-view radar representations, whereas RC-ROSNet [14] additionally incorporates sequential visual images alongside multi-frame radar data. T-RODNet [15] also uses multi-frame radar inputs but adopts a Transformer-based architecture for spatio-temporal feature modeling. In contrast, TransRadar [16] operates on single-frame radar inputs. Collectively, these methods provide representative baselines spanning different input configurations and feature modeling strategies. In addition, we evaluated several conventional methods, including CA-CFAR [17], OS-CFAR [17], GO-CFAR [17], SO-CFAR [17], and PCA-CFAR [18]. To improve their capability to detect weak targets, coherent integration was performed prior to CFAR detection.

In radar target detection, the widely used object location similarity (OLS) [1] is employed as a localization similarity measure, analogous to intersection over union (IoU) in image-based object detection. The corresponding calculation formula is given below:

$$\mathrm{OLS} = \exp\left\{\frac{-d^2}{2 \times (s \times k_{\mathrm{cls}})^2}\right\} \tag{C1}$$

where d denotes the distance between two points in the confidence map, s denotes the distance between the target and the radar sensor, and k_{cls} denotes the class-specific tolerance coefficient, which is determined by the average size of targets in each category. During evaluation, the OLS between each detection and the corresponding ground-truth annotation is first computed for every frame. Based on predefined OLS thresholds, mean average precision (mAP) and mean average recall (mAR) are then calculated. To comprehensively evaluate model performance, we employ four metrics: mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$. Specifically, mAP_{50} reflects the model's basic target recognition capability, mAP_{70} indicates its localization accuracy, $\mathrm{mAP}_{50:90}$ provides an overall assessment by averaging detection performance across multiple stricter thresholds, and $\mathrm{mAR}_{50:90}$ measures the model's ability to cover ground-truth targets, thereby effectively indicating its missed-detection tendency.

Appendix C.1.2 *Implementation details*

We selected RODNet-CDC as the baseline model. All comparative experiments were conducted on a platform equipped with an NVIDIA RTX 5000 Ada Generation GPU. During training, all models were optimized using the AdamW optimizer, with an initial learning rate of 1×10^{-4}. The learning rate was updated on a per-batch basis according to a cosine annealing schedule, with a minimum value of 1×10^{-6}. The entire training procedure lasted 55 epochs. To alleviate potential neuron inactivation caused by excessively large gradients during the early stage of training, a linear warm-up strategy was employed in the first 5 epochs, during which the learning rate was gradually increased from 1×10^{-7} to 1×10^{-4}. Over the subsequent 50 epochs, the learning rate was progressively decreased following the cosine annealing schedule to promote stable convergence. In addition, unless otherwise specified, all methods adopted a sliding-window strategy with a stride of 4, and this setting was kept consistent across both the training and validation stages to ensure alignment between the model input format and the evaluation pipeline, thereby reducing potential bias introduced by inconsistent stride settings. To ensure reproducibility, the random seed was set to 2027 for all experiments. Notably, TransRadar and RC-ROSNet were specifically designed for single-frame radar object detection; accordingly, their sliding-window size was set to 1, and the number of input radar frames, F, remained fixed at 1 throughout both training and inference.

During training, the input data are represented as a tensor of size $B \times C_{\mathrm{input}} \times N_{\mathrm{chirp}} \times F \times R \times A$, where B denotes the batch size, which is set to 4, and C_{input} denotes the number of input channels, which is set to 2 to represent the real and imaginary components of

Table C1 Comparison of STCT-Net with SOTA radar object detection methods on UAVRadar. The best results are in bold. The underlined values indicate the best results among the compared methods for each metric. "*" means the use of the PCA-based sampling strategy.

Methods	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)	Parameters (M)	GFLOPs	FPS
CA-CFAR	7.85	5.21	4.70	17.01	-	-	-
OS-CFAR	7.14	5.02	4.55	18.86	-	-	-
GO-CFAR	7.56	4.32	4.22	15.06	-	-	-
SO-CFAR	6.93	4.51	4.18	19.13	-	-	-
PCA-CFAR	5.82	2.55	2.71	7.71	-	-	-
RODNet-CDC	80.27	69.93	64.20	70.64	34.93	390.47	682.76
RODNet-CDC*	85.09	73.17	67.70	76.12	34.93	545.66	55.52
RODNet-HG	82.36	64.86	59.80	67.36	64.81	1130.70	243.20
RODNet-HG*	84.92	68.89	65.36	72.93	64.81	1275.89	46.95
RODNet-HGwI	84.48	58.93	56.05	67.78	61.14	5939.92	54.01
RODNet-HGwI*	86.30	69.28	65.81	73.45	61.14	6085.11	29.85
T-RODNet	72.87	55.40	50.76	61.99	31.54	175.05	433.29
T-RODNet*	66.42	51.99	49.67	57.92	31.54	320.24	51.89
RAMP-CNN	60.92	45.81	44.21	51.72	104.21	1458.34	177.48
E-RODNet	80.40	65.29	58.42	67.68	6.10	**36.03**	151.77
E-RODNet*	85.84	68.84	63.46	71.90	6.10	181.22	48.17
TransRadar	52.95	40.05	38.66	50.54	**5.12**	68.14	**780.38**
TransRadar*	57.68	44.55	42.27	52.27	**5.12**	213.33	313.76
RC-ROSNet	75.56	60.43	56.07	64.29	47.28	268.83	77.19
RC-ROSNet*	68.79	54.41	51.10	59.94	47.28	414.02	36.55
STCT-Net (ours)	87.09	75.79	72.20	76.95	6.32	46.01	357.74
STCT-Net* (ours)	**91.16**	**82.41**	**76.55**	**81.14**	6.32	191.20	53.26

the input data. N_{chirp} denotes the number of chirps in each radar frame and, unless otherwise specified, is set to 32 under a uniform sampling strategy; a more detailed analysis of this parameter is provided later. F denotes the length of the input frame sequence and is set to 16. R and A denote the height and width of the input data, corresponding to the numbers of range bins and angle bins, respectively, both of which are set to 128. No image-based data augmentation was applied during training, because radar signal processing differs fundamentally from image processing, and conventional image augmentation methods are generally unsuitable for radar data. In addition, RAMP-CNN and TransRadar take additional RD and AD spectra as inputs. The RD spectrum has an input dimension of $B \times C_{\text{input}} \times F \times R \times D$, where D denotes the width of the input data and is set to 128, corresponding to the number of Doppler bins. The AD spectrum has an input dimension of $B \times C_{\text{input}} \times F \times A \times D$. Furthermore, RC-ROSNet incorporates additional image inputs to enable cross-modal training and inference. In contrast, T-RODNet is designed to perform prediction using only a single chirp as input; therefore, its N_{chirp} is fixed at 1.

Appendix C.2 Comparison with SOTA methods

We systematically evaluated the detection performance of the proposed STCT-Net against other SOTA models and provided a detailed experimental report. Table C1 summarizes the evaluation metrics for the UAV radar target detection task on the UAVRadar benchmark, including mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$. The results show that STCT-Net achieves 91.16% mAP_{50}, 82.41% mAP_{70}, 76.55% $mAP_{50:90}$, and 81.14% $mAR_{50:90}$, outperforming other SOTA methods across all metrics and demonstrating a substantial performance advantage. To ensure a fair comparison among different methods, we further evaluated all models under a unified setting. Specifically, all methods were tested under identical experimental conditions. Methods marked with "*" denote the use of the PCA-based sampling strategy introduced in this study, whereas methods without an asterisk use their default uniform sampling strategy. Under this unified configuration, STCT-Net still achieves 72.20% $mAP_{50:90}$ and 76.95% $mAR_{50:90}$. Compared with the baseline model, its $mAP_{50:90}$ and $mAR_{50:90}$ improve by 13.95 and 11.63 percentage points, respectively. These results indicate that STCT-Net not only delivers outstanding overall detection accuracy but also maintains a clear performance advantage under a unified sampling strategy, thereby confirming its effectiveness and robustness in UAV radar target detection. They further suggest that the network architecture and feature-modeling strategy proposed in this study can exploit radar spatio-temporal information more effectively, thereby enhancing target detection accuracy and recall. To assess the applicability of the PCA sampling strategy across input configurations, we compared each method's detection performance with and without the strategy. The multi-frame, multi-chirp methods RODNet-CDC, RODNet-HG, RODNet-HGwI, and E-RODNet showed varying degrees of improvement. In contrast, performance declined for the multi-frame, single-chirp method T-RODNet and the multimodal method RC-ROSNet, which uses the same radar input configuration as T-RODNet. For the single-frame, single-chirp method TransRadar, $mAP_{50:90}$ increased from 42.27% to 56.07%, a gain of 13.80 percentage points. These findings indicate that the effectiveness of the PCA strategy varies across methods and cannot be determined solely by the number of input frames or chirps. We hypothesize that these differences are related to the amount of target information retained after compression and the models' feature extraction and fusion mechanisms. When few samples are retained, PCA compression may discard useful information about small, weak targets, thereby impairing detection performance. For single-frame models, compression along the micro-time dimension may improve target representations. For multi-frame models, cross-frame fusion may provide complementary target information that is difficult to discern in individual frames. However, effective use of compressed inputs may still depend on

Table C2 Ablation results of different components in STCT-Net.

Method	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
Full model	91.16	82.41	76.55	81.14
w/o PCA-based sampling	87.09(↓ 4.07)	75.79(↓ 6.62)	72.20(↓ 4.35)	76.95(↓ 4.19)
w/o STCT module	90.23(↓ 0.93)	76.86(↓ 5.55)	72.44(↓ 4.11)	78.78(↓ 2.36)
w/o optimized feature extractor	88.31(↓ 2.85)	74.05(↓ 8.36)	69.17(↓ 7.38)	76.02(↓ 5.12)
w/o Soft-BCE loss	89.66(↓ 1.50)	80.33(↓ 2.08)	74.78(↓ 1.77)	80.76(↓ 0.38)

the specific feature extraction and fusion architecture. Therefore, the applicability of the PCA strategy should be assessed by jointly considering the input configuration and network architecture. The proposed mechanisms require further validation.

In addition, we compared the inference cost and parameter count of various radar target detection models using GFLOPs, the number of parameters, and frames per second (FPS). The results show that the proposed STCT-Net requires only 10.01 additional GFLOPs and 0.23 million more parameters than the lowest-FLOPs model, E-RODNet, while achieving superior performance. Moreover, by employing a feature encoder based on depthwise convolutions, the proposed method is lighter than the baseline model and delivers better performance. When the computational cost of PCA preprocessing is included, the overall computational complexity increases from 46.01 to 191.20 GFLOPs, while the inference speed decreases from 357.74 to 53.26 FPS, indicating that the preprocessing step incurs substantial additional computational overhead. Despite this overhead, when considering the network architecture itself, STCT-Net maintains a favorable trade-off between computational complexity and detection performance, highlighting its computational efficiency relative to the compared methods. These findings demonstrate that the proposed method achieves superior performance at the cost of only a modest increase in inference overhead.

Appendix C.3 Ablation study

In this subsection, we conduct ablation studies to verify the effectiveness of each component in STCT-Net.

Appendix C.3.1 *Module analysis*

We present an ablation study of the key components of the proposed STCT-Net in Table C2, including the PCA-based sampling strategy, the STCT module, the optimized feature extractor, and the Soft-BCE loss function. Specifically, we used the complete STCT-Net as the reference model and conducted comparative experiments by removing each key component individually to evaluate its contribution to the final detection performance more intuitively. For each removed component, the corresponding module from the baseline model, RODNet-CDC, was used as a replacement to ensure consistency in the overall network architecture and experimental settings. The experimental results show that the full model achieves the best performance across all evaluation metrics, indicating that these modules produce strong synergistic effects and jointly enhance detection performance. When the PCA-based sampling strategy is removed, $mAP_{50:90}$ and $mAR_{50:90}$ decrease by 4.35 and 4.19 percentage points, respectively, demonstrating that this strategy effectively preserves critical dynamic information for small and weak targets and provides more discriminative input representations for the subsequent network. When the STCT module is removed, performance also declines markedly, especially on the stricter metric mAP_{70} and the overall metric $mAP_{50:90}$, which decrease by 5.55 and 4.11 percentage points, respectively. This result indicates that the proposed chirp merging mechanism effectively enhances the model's ability to jointly capture target dynamic and spatial structural features. Furthermore, replacing the optimized feature extractor proposed in this study with the baseline feature extractor leads to the most substantial performance degradation, with $mAP_{50:90}$ and mAP_{70} decreasing by 7.38 and 8.36 percentage points, respectively. This finding suggests that the enhanced three-dimensional modeling capability of the feature extractor, together with its structural optimization tailored to the characteristics of radar data, plays a decisive role in improving overall detection performance. In addition, removing the Soft-BCE loss function leads to declines in all performance metrics to varying degrees, indicating that this loss function is better suited to Gaussian heatmap supervision, enhances the modeling of target-center responses, and effectively suppresses background noise. Overall, these ablation experiments validate the effectiveness of each component in the proposed framework. Among them, the optimized feature extractor contributes most substantially to the performance improvement, while the PCA-based sampling strategy, the STCT module, and the Soft-BCE loss function also play important roles. These findings indicate that the performance gains of the proposed method arise not from the improvement of any single module, but from the synergistic optimization of multiple components, thereby significantly enhancing the detection capability for small and weak UAV targets.

Appendix C.3.2 *Impact of uniform and PCA-based sampling under different sample numbers*

To investigate the effects of different sampling strategies and numbers of sampled chirps on the performance of radar object detection networks, we conducted a series of comparative experiments. Specifically, we compared the proposed PCA-based sampling strategy with the uniform sampling strategy commonly used in radar object detection. To ensure a fair evaluation, both strategies were assessed using the same numbers of sampled chirps. In particular, we evaluated model performance with 4, 8, 16, 32, and 64 chirps retained by uniform sampling, as well as with 4, 8, 16, 32, and 64 chirps retained by PCA-based sampling. Among these settings, uniform sampling with 4 chirps is the most commonly used configuration in radar object detection and was therefore adopted as the baseline for comparison. As shown in Table C3, under their respective sampling strategies, both uniform sampling with 32 chirps and PCA-based sampling with 32 chirps achieved the best $mAP_{50:90}$ within each setting. Compared with the baseline, uniform sampling with 32 chirps improved mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ by 2.98, 18.01, 14.23, and 8.39 percentage points, respectively. These results indicate that, for weak UAV radar target detection, increasing the number of sampled chirps helps improve detection performance. Therefore, in the comparative experiments reported in Table C1, all competing methods that support multi-chirp input were evaluated

Table C3 Ablation results comparing PCA-based and uniform sampling strategies under different numbers of sampled chirps. The best results are in bold.

Number of Chirps	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
Uniform				
4	84.11	57.78	57.97	68.56
8	83.84	67.41	64.71	72.64
16	81.20	65.75	63.39	69.30
32	**87.09**	**75.79**	**72.20**	**76.95**
64	82.32	67.13	64.89	71.15
PCA				
4	78.35	64.82	61.22	69.12
8	81.39	68.40	66.10	72.38
16	85.46	74.02	69.86	75.79
32	91.16	**82.41**	**76.55**	**81.14**
64	**92.10**	81.97	76.13	81.07

Table C4 Experimental results comparing PCA-based and uniform sampling under their respective optimal numbers of sampled chirps in low- and high-SNR scenarios.

Sampling Strategy	**SNR Condition**	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
Uniform	SNR $<$ 2.23	78.89	74.21	69.65	71.74
PCA-based	SNR $<$ 2.23	86.99(↑ 8.10)	82.04(↑ 7.83)	73.76(↑ 4.11)	78.55(↑ 6.81)
Uniform	SNR $\geqslant$ 2.23	86.03	74.49	70.92	75.95
PCA-based	SNR $\geqslant$ 2.23	91.19(↑ 5.16)	81.39(↑ 6.90)	76.62(↑ 5.70)	80.86(↑ 4.91)

using a uniform-sampling configuration with 32 chirps rather than 4 chirps, thereby ensuring consistency and fairness in the experimental settings.

The PCA-based sampling strategy achieves optimal performance when 32 chirps are retained. Compared with the baseline model, it improves mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ by 7.05, 24.63, 18.58, and 12.58 percentage points, respectively. However, when only 4 chirps are retained, the mAP_{50} drops markedly to 78.35%. This degradation may be attributed to the weak reflectivity of UAV targets in urban environments. When the dimensionality after PCA-based reduction is too low, the retained information is more likely to be dominated by noise and therefore cannot adequately represent the key characteristics of the targets, resulting in inferior detection performance. Nevertheless, the PCA-based sampling strategy still exhibits certain advantages, as it continues to outperform the baseline in terms of mAP_{70}, the comprehensive metric $mAP_{50:90}$, and the recall metric $mAR_{50:90}$. In addition, we compared the SNR distributions of data obtained with different numbers of PCA-sampled chirps with that of the original data. As shown in Figure C1, as the number of sampled chirps increases, the SNR distribution of the downsampled data gradually approaches that of the original data. When the SNR of the original data is below 0 dB, excessively small sampling numbers, as illustrated in Figure C1(a) and Figure C1(b), significantly increase the proportion of negative-SNR targets, indicating that noise is amplified while the target signal is weakened. This observation also explains the performance degradation when the number of PCA-sampled chirps is too small. Essentially, the PCA-based sampling strategy improves the SNR of low-SNR samples while maintaining a relatively stable proportion of negative-SNR samples. For example, a comparison between Figure C1(e) and Figure C1(f) shows that the number of samples near the peak (SNR $\approx$ 0 dB) decreases by approximately 10%, whereas the number of negative-SNR samples remains nearly unchanged, demonstrating an advantage over the distribution shown in Figure C1(d). Moreover, once the number of sampled chirps reaches an appropriate level, high-SNR targets become more concentrated in the 2-10 dB range. We further estimated the SNR distribution of targets in the original dataset through coherent integration, and the results show that the average SNR of UAV targets is approximately 2.23 dB. This finding further confirms that target signals are more concentrated in the 2-10 dB interval.

To further analyze the effect of PCA-based sampling on the SNR distribution, we examined it from the perspective of information retention. Here, the information retention rate was calculated from the RA spectrum sequence of the entire radar frame. When PCA-based sampling retains only one chirp, 4.37% of the information is preserved. As the number of sampled chirps increases to 4, 8, 16, 32, and 64, the information retention rate rises to 8.61%, 12.47%, 18.94%, 30.01%, and 47.79%, respectively. Combined with the preceding experimental results, these findings indicate that the model achieves the best performance when PCA-based sampling retains 32 chirps. Given that the average SNR of the original data is 2.23 dB and that most samples are distributed around approximately 0 dB, we attribute this result primarily to two factors. First, this setting achieves a favorable balance between preserving critical information and suppressing noise. Compared with retaining 4, 8, or 16 chirps, it preserves more target-related information and thus reduces feature loss. In contrast, compared with retaining 64 chirps, it avoids introducing excessive redundant information and noise interference. In addition, based on the average SNR, we further compared the uniform sampling strategy and the PCA-based sampling strategy under their respective optimal configurations. Specifically, using the average SNR as the criterion, we divided the experimental scenarios into low-SNR and high-SNR conditions. Low-SNR scenarios correspond to SNR $<$ 2.23 and mainly include hovering, lateral motion, or long-range targets, whereas high-SNR scenarios correspond to SNR $\geqslant$ 2.23 and mainly include longitudinal motion or short-range targets. The results in Table C4 show that the PCA-based sampling strategy outperforms the uniform sampling strategy in both types of scenarios, with a more pronounced advantage under low-SNR conditions. Specifically, in low-SNR scenarios, mAP_{50}, mAP_{70}, $mAP_{50:90}$,

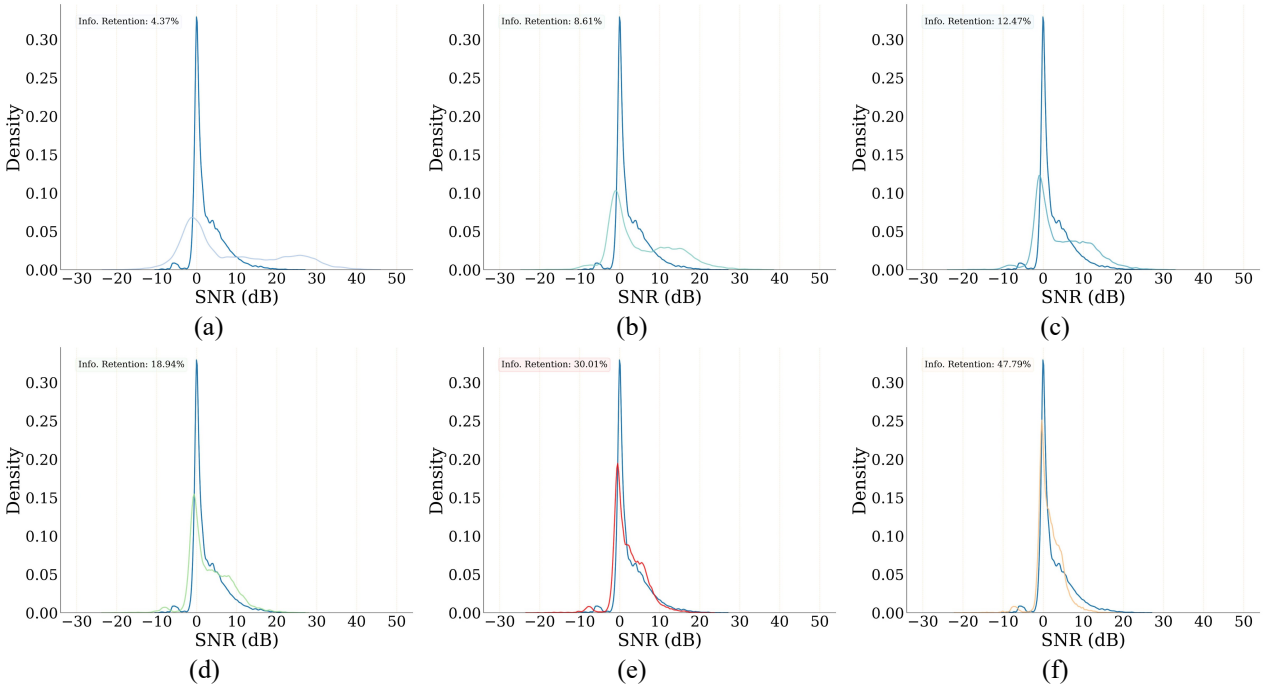

Figure C1 Kernel density estimation of the SNR distribution, expressed in dB, is used to evaluate the effect of PCA-based sampling on the data distribution. We compare the distributions of the original dataset and the PCA-reduced dataset under different numbers of sampled chirps: (a) 1 chirp, (b) 4 chirps, (c) 8 chirps, (d) 16 chirps, (e) 32 chirps, and (f) 64 chirps.

Table C5 Ablation results for different components of the STCT module. The best results are in bold.

Spatial branch	Temporal branch	Branch fusion unit	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
✓	✗	✗	87.32	76.05	70.66	76.16
✗	✓	✗	71.83	55.47	59.33	51.77
✓	✓	✗	86.94	72.00	68.61	74.28
✓	✓	✓	**91.16**	**82.41**	**76.55**	**81.14**

and $mAR_{50:90}$ improve by 8.10, 7.83, 4.11, and 6.81 percentage points, respectively. In particular, $mAR_{50:90}$ increases from 71.74% to 78.55%, indicating that the PCA-based sampling method can preserve target-related salient features more effectively under low-SNR conditions and mitigate the impact of noise interference on missed detections, thereby significantly improving the model's detection capability. Under high-SNR conditions, PCA-based sampling also yields consistent performance gains, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ improving by 5.16, 6.90, 5.70, and 4.91 percentage points, respectively. Although the original echoes already exhibit good separability under high-SNR conditions, PCA-based sampling can further enhance target feature representation by preferentially preserving principal-component information while suppressing secondary redundant components. Overall, the PCA-based sampling strategy not only achieves better overall performance than the uniform sampling strategy but also exhibits greater robustness and more substantial performance gains under low-SNR conditions. These results demonstrate that this strategy can effectively select critical temporal information in complex radar sensing environments, thereby improving UAV radar target detection performance.

Appendix C.3.3 *The effectiveness of STCT module*

To validate the effectiveness of the proposed STCT module within the overall framework, we conducted two groups of ablation experiments to systematically evaluate the contributions of its constituent components to detection performance.

The first group focused on the three core components of the STCT module, namely the spatial branch, the temporal branch, and the branch fusion unit, with the aim of analyzing their individual roles and synergistic effects in spatio-temporal feature modeling. As shown in Table C5, when only the spatial branch is included, the model achieves 87.32% mAP_{50}, 76.05% mAP_{70}, 70.66% $mAP_{50:90}$, and 76.16% $mAR_{50:90}$, indicating relatively strong overall performance. This result suggests that the spatial branch can effectively extract spatial structural features and thus provides substantial support for detection performance. In contrast, when only the temporal branch is included, performance drops markedly, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reaching only 71.83%, 55.47%, 59.33%, and 51.77%, respectively. This finding indicates that, in the absence of spatial structural constraints, temporal information alone is insufficient to form effective target representations. A possible explanation is that adjacent bins in the RA spectrum exhibit strong local correlations; if spatial coupling relationships are ignored, modeling only temporal features may disrupt spatial continuity, thereby weakening overall detection capability. When both the spatial and temporal branches are included without the branch fusion unit, model performance does not improve further, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reaching 86.94%, 72.00%, 68.61%, and 74.28%, respectively. Compared with the configuration using only the spatial branch, this setting instead yields slight decreases

Table C6 Comparison of temporal compression strategies within the spatial branch of the STCT module. The best results are in bold.

Max	Average	Mixed	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
✓			86.30	74.88	69.51	75.94
	✓		**91.16**	**82.41**	**76.55**	**81.14**
		✓	86.52	76.47	69.41	76.34

Table C7 Comparison of different chirp merging methods. The best results are in bold.

Chirp merging method	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
MNet	90.23	76.86	72.44	78.78
SSFM	86.88	72.81	67.90	75.69
STCT module (ours)	**91.16**	**82.41**	**76.55**	**81.14**

Table C8 Ablation results of key components in the optimized feature extractor. The best results are in bold.

MLP to Conv	LN to GN	StarReLU to GELU	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
✗	✗	✗	89.55	75.71	70.10	78.22
✓	✗	✗	84.90	72.59	68.15	74.65
✗	✓	✗	83.91	63.53	63.15	70.67
✗	✗	✓	84.80	68.15	65.07	74.22
✓	✓	✗	75.81	59.47	57.43	64.48
✓	✗	✓	84.35	69.74	65.35	72.71
✗	✓	✓	90.12	79.90	74.56	79.64
✓	✓	✓	**91.16**	**82.41**	**76.55**	**81.14**

across all metrics, indicating that merely introducing spatial and temporal information through a parallel structure cannot fully exploit their complementary advantages. Without an effective mechanism for feature interaction and fusion, the model's capacity for joint spatio-temporal modeling remains limited. When the spatial branch, temporal branch, and branch fusion unit are all included, the model achieves the best performance, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reaching 91.16%, 82.41%, 76.55%, and 81.14%, respectively. Compared with the other configurations, the complete STCT module performs best across all evaluation metrics, with particularly notable advantages on the stricter metric mAP_{70} and the comprehensive metric $mAP_{50:90}$. These findings indicate that the branch fusion unit plays a critical role in spatio-temporal feature interaction by effectively integrating the complementary information extracted by the spatial and temporal branches, thereby significantly enhancing the model's ability to jointly capture target dynamics and spatial structure. Overall, the spatial branch provides the foundation for detection performance, whereas the temporal branch alone is insufficient for effective detection. Only with the support of the branch fusion unit can the synergistic advantages of the spatial and temporal branches be fully realized. These results validate the soundness of the overall STCT module design and its effectiveness in improving the detection performance of small and weak targets.

The second group of ablation experiments examined the effect of different temporal compression strategies in the spatial branch on the performance of the STCT module. Because the STCT module already includes an independent temporal branch for modeling temporal features, the spatial branch must compress the temporal dimension through pooling so that its output dimensionality matches that of the temporal branch. In this study, we evaluated three temporal compression strategies: max pooling, average pooling, and hybrid pooling. In the hybrid pooling scheme, temporal information is first compressed separately by max pooling and average pooling, and the final output of the spatial branch is then generated through weighted fusion. As shown in Table C6, different temporal compression strategies have a significant impact on model performance. Among them, average pooling achieves the best results, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reaching 91.16%, 82.41%, 76.55%, and 81.14%, respectively. In contrast, max pooling performs substantially worse, with the corresponding metrics reaching 86.30%, 74.88%, 69.51%, and 75.94%, respectively. Hybrid pooling yields 86.52%, 76.47%, 69.41%, and 76.34%, respectively, and its overall performance is likewise clearly inferior to that of average pooling. Although hybrid pooling performs slightly better than max pooling in terms of mAP_{50}, mAP_{70}, and $mAR_{50:90}$, its $mAP_{50:90}$ is slightly lower than that of max pooling, indicating that it does not provide a stable advantage in overall detection performance. In addition, the weak-target radar detection task considered in this study is characterized by low signal strength and high susceptibility to noise interference. Max pooling tends to preserve local maximum responses along the temporal dimension, but in weak-target scenarios, these instantaneous high values often originate from noise or clutter, thereby weakening the effective representation of true target features. By contrast, average pooling aggregates features more smoothly along the temporal dimension, suppressing random noise while better preserving the overall energy distribution of the target signal. It is therefore more suitable for weak-target radar sensing tasks. Although hybrid pooling attempts to combine the advantages of max pooling and average pooling, its fusion process may introduce additional feature perturbations and thus fails to further improve model performance. Overall, among the three evaluated temporal compression strategies, average pooling is the most suitable for temporal compression in the spatial branch of the STCT module. It not only achieves the best results across all evaluation metrics but also exhibits more stable and effective feature-modeling capability for weak targets.

Finally, we compared several chirp-merging methods, including MNet and the SSFM, and the results are presented in Table C7. Compared with the other methods, the proposed STCT module achieves the best performance across all evaluation metrics, indicating that the joint modeling of spatial and temporal features preserves input information more effectively and thereby improves target detection performance.

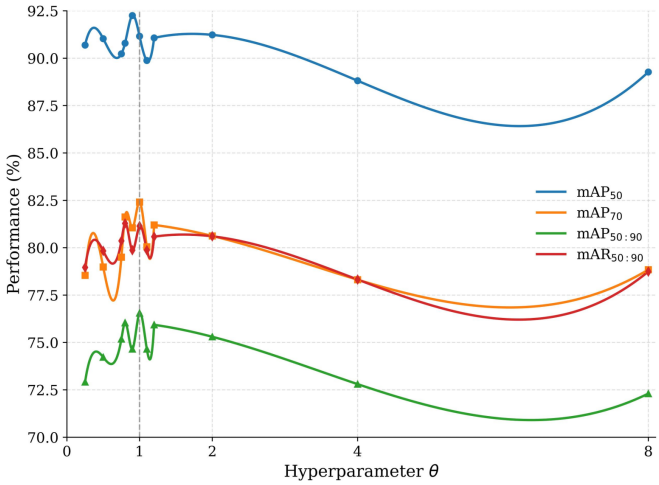

Figure C2 Ablation results of the hyperparameter θ in the proposed Soft-BCE loss.

Appendix C.3.4 *The impact of feature extractor component replacement*

To evaluate the impact of key structural components in the backbone network on model performance, we conducted ablation experiments on the core design choices of the optimized feature extractor. Specifically, we examined three architectural modifications: replacing the projection layer in the original backbone from an MLP with a $1 \times 1 \times 1$ convolution (denoted as MLP to Conv), replacing LN with GN, and replacing the StarReLU activation function with GELU. The results, presented in Table C8, reveal substantial differences in how these structural components affect model performance. When only MLP to Conv is applied, performance declines relative to the baseline, with mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$ decreasing from 89.55%, 75.71%, 70.10%, and 78.22% to 84.90%, 72.59%, 68.15%, and 74.65%, respectively. This indicates that replacing the MLP with a convolution alone does not yield performance gains and that its effectiveness depends on synergy with other architectural modifications. Similarly, when only LN is replaced with GN, performance degrades even more markedly, with the corresponding metrics dropping to 83.91%, 63.53%, 63.15%, and 70.67%, respectively. This suggests that, without other coordinated optimizations, GN cannot adequately adapt to the feature-distribution characteristics of the current task and therefore adversely affects detection performance. By contrast, replacing only StarReLU with GELU leads to a relatively smaller performance drop, with mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$ reaching 84.80%, 68.15%, 65.07%, and 74.22%, respectively. This result indicates that modifying the activation function alone is likewise insufficient to improve overall performance. Furthermore, under pairwise combinations, different components exhibit markedly different synergistic effects. When MLP to Conv and LN to GN are applied simultaneously, the model suffers the most severe performance degradation, with mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$ reaching only 75.81%, 59.47%, 57.43%, and 64.48%, respectively. This indicates that, in the absence of a suitable activation function, these two modifications may have substantial negative effects on feature representation and training stability. When MLP to Conv is combined with GELU, the model still performs worse than the baseline, suggesting that convolution replacement and activation adjustment alone are insufficient to produce effective complementarity. Notably, when LN to GN and StarReLU to GELU are adopted simultaneously, model performance improves significantly, with mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$ reaching 90.12%, 79.90%, 74.56%, and 79.64%, respectively, all exceeding the baseline. This indicates a strong synergy between the normalization scheme and the activation function: the smoother nonlinear behavior of GELU can, to some extent, alleviate the feature-distribution fluctuations introduced by GN, thereby enhancing overall feature-modeling capability. When all three modifications are introduced simultaneously, the model achieves the best performance, with mAP_{50}, mAP_{70}, $\mathrm{mAP}_{50:90}$, and $\mathrm{mAR}_{50:90}$ reaching 91.16%, 82.41%, 76.55%, and 81.14%, respectively. Compared with the baseline model, these values represent improvements of 1.61, 6.70, 6.45, and 2.92 percentage points, respectively. In particular, the gains on the stricter metric mAP_{70} and the comprehensive metric $\mathrm{mAP}_{50:90}$ are more pronounced, indicating that the joint introduction of multiple structural optimizations can effectively enhance feature representation and model robustness.

Appendix C.3.5 *The sensitivity analysis of Soft-BCE loss hyperparameter*

To evaluate the effectiveness of the proposed Soft-BCE loss function in weak UAV radar target detection, we conducted experiments from two perspectives. First, we performed a sensitivity analysis of the key hyperparameter θ in Soft-BCE. Second, we compared Soft-BCE with several commonly used loss functions, including BCE, MAE, MSE, smooth-L1, focal loss, and penalty-reduced focal loss.

To analyze the effect of the hyperparameter θ on model performance, we adopted a two-stage search strategy, progressing from coarse- to fine-grained exploration. In the coarse-grained stage, $\theta = 0.25, 0.5, 0.75, 1, 2, 4$, and 8 were selected for preliminary evaluation. Based on these results, a fine-grained search was subsequently conducted around $\theta = 1$, with $\theta = 0.8, 0.9, 1.1$, and 1.2 selected for further assessment within the better-performing interval identified in the coarse search. The experimental results are shown in Figure C2. Overall, as θ increases from smaller values, model performance first improves and then declines, indicating that this hyperparameter has a significant effect on both the optimization behavior of the loss function and the final detection performance. Notably, as θ approaches 0, the soft adjustment effect of Soft-BCE gradually weakens, and the loss correspondingly degenerates into the standard BCE loss. In other words, under this condition, the loss function distinguishes positive and negative samples in the same manner as conventional BCE and therefore cannot exploit the advantages of soft-label modeling. Accordingly, a moderate increase in θ enhances the ability of the

Table C9 Comparison of different loss function. The best results are in bold.

Loss function	mAP_{50} (%)	mAP_{70} (%)	$mAP_{50:90}$ (%)	$mAR_{50:90}$ (%)
BCE	89.66	80.33	74.78	80.76
MAE	87.18	79.56	73.23	78.53
MSE	85.76	76.23	72.02	76.80
Smooth-L1	88.32	76.90	71.82	78.52
Focal loss	89.24	74.78	71.01	77.13
Penalty-reduced focal loss	84.21	71.96	68.09	74.10
Soft-BCE (ours)	**91.16**	**82.41**	**76.55**	**81.14**

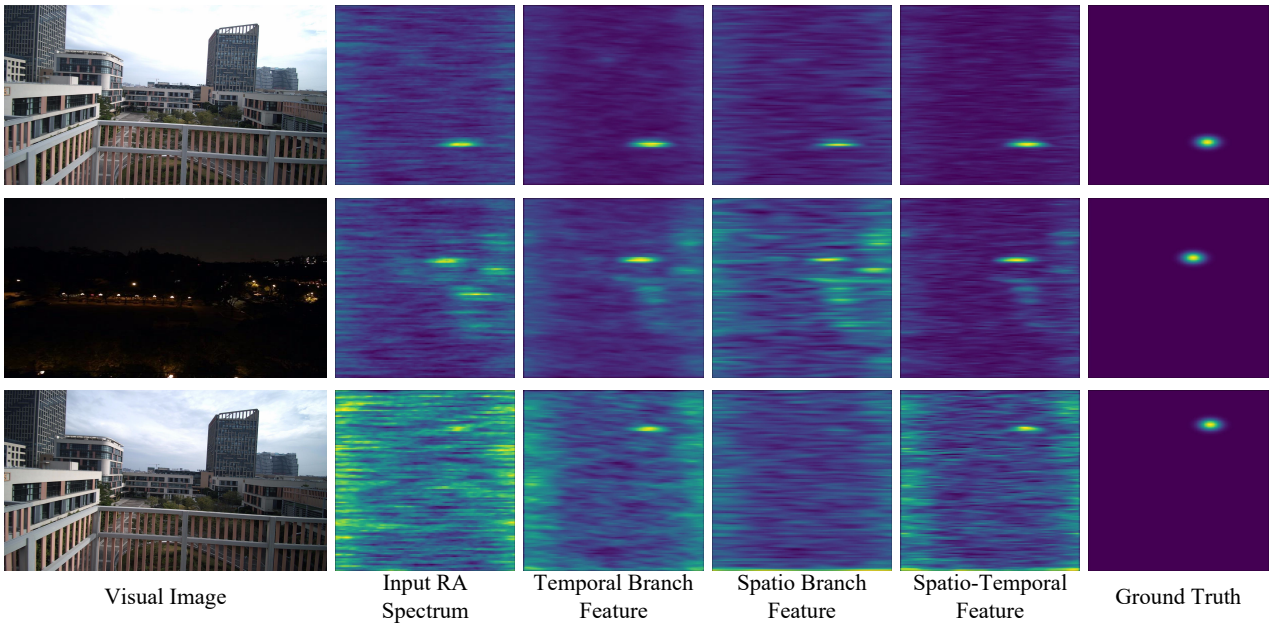

Figure C3 Visualization of feature representations from different branches of the STCT module.

loss function to characterize the continuous differences between targets and background, thereby improving the model's discriminative capability for small and weak targets. Specifically, when $\theta = 0.25$, the model achieves mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ of 90.69%, 78.54%, 72.91%, and 78.96%, respectively. As θ increases to 0.5 and 0.75, the overall model performance improves progressively, with $mAP_{50:90}$ rising to 74.22% and 75.18%, respectively. Furthermore, when θ lies approximately in the range of 0.8 to 1.0, the model reaches a relatively favorable performance level. In particular, when $\theta = 1.0$, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reach 82.41%, 76.55%, and 81.14%, respectively, yielding the best overall detection performance. By contrast, when $\theta = 0.9$, the model attains the highest mAP_{50} of 92.25%, indicating that even small variations in this parameter may affect different evaluation metrics differently. Considering performance across all evaluation metrics, we determined the value of θ primarily based on mAP_{70}, which reflects localization accuracy, and $mAP_{50:90}$, which reflects overall detection performance, and ultimately set it to 1.0. However, when θ is further increased to 2, 4, and 8, model performance declines to varying degrees. In particular, when $\theta = 4$ and $\theta = 8$, $mAP_{50:90}$ decreases to 72.80% and 72.30%, respectively. These results indicate that an excessively large θ weakens the ability of the loss function to effectively model soft-label differences between targets and background, thereby impairing final detection performance. In summary, the value of θ has a significant effect on the performance of Soft-BCE: when it is too small, the loss function gradually degenerates into BCE and cannot fully exploit the advantages of soft-label modeling; when it is too large, it weakens the ability to characterize fine-grained inter-class differences.

Furthermore, we compared Soft-BCE with several commonly used loss functions, and the experimental results are presented in Table C9. The proposed Soft-BCE achieves the best performance across all evaluation metrics, with mAP_{50}, mAP_{70}, $mAP_{50:90}$, and $mAR_{50:90}$ reaching 91.16%, 82.41%, 76.55%, and 81.14%, respectively. By contrast, conventional BCE yields the second-best performance, with the corresponding metrics reaching 89.66%, 80.33%, 74.78%, and 80.76%, respectively. MAE and MSE exhibit relatively weaker overall performance, with MSE achieving only 85.76%, 76.23%, 72.02%, and 76.80% on the four metrics, respectively. Although smooth-L1 performs slightly better than MSE, it remains clearly inferior to Soft-BCE in terms of the overall metrics. For the two loss functions that emphasize hard-sample learning, namely focal loss and penalty-reduced focal loss, performance is likewise inferior to that of the proposed method. In particular, penalty-reduced focal Loss performs the worst, with an $mAP_{50:90}$ of only 68.09%. These results indicate that Soft-BCE is better suited to weak UAV radar target detection. The main reason is that this loss function is more compatible with Gaussian heatmap supervision. While preserving the discriminative properties of BCE for classification, it imposes finer constraints on the foreground region through softened modeling of target responses, thereby enhancing the model's ability to represent the target center and its neighboring regions. Compared with MAE, MSE, and smooth-L1, which directly treat the heatmap as a numerical regression target, Soft-BCE can more effectively distinguish target regions from the background. In contrast to focal Loss [19] and its variants [13], Soft-BCE avoids the training instability caused by excessive emphasis on a small number of hard samples,

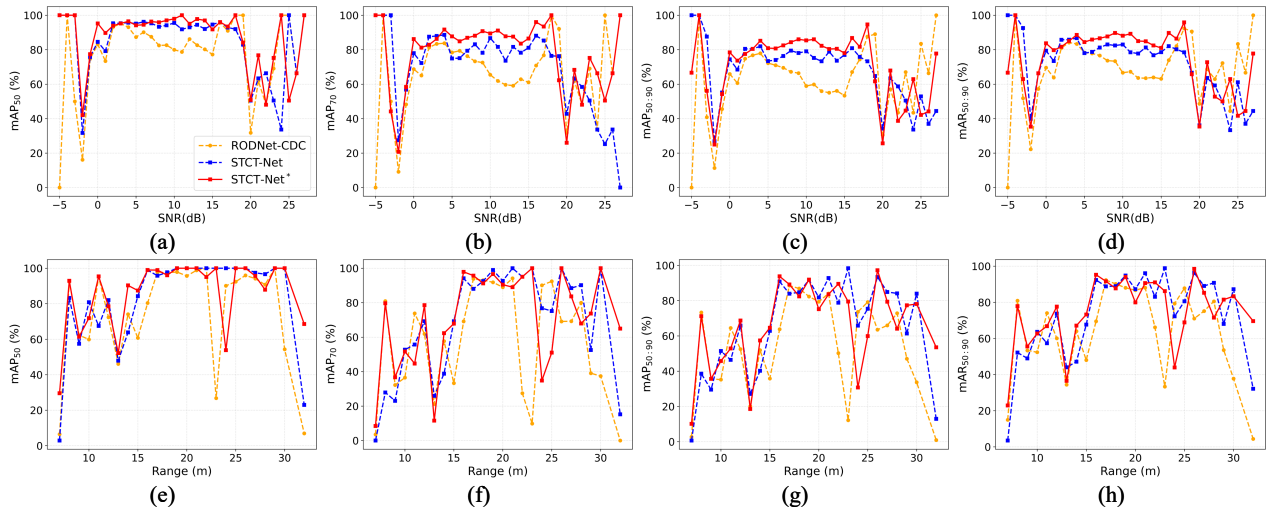

Figure C4 Comparison of the proposed STCT-Net and STCT-Net* with the baseline model RODNet-CDC under different SNR and target-range conditions. (a)-(d) detection performance as a function of SNR, evaluated using mAP_{50}, mAP_{70}, $\text{mAP}_{50:90}$, and $\text{mAR}_{50:90}$, respectively. (e)-(h) corresponding detection performance as a function of target range.

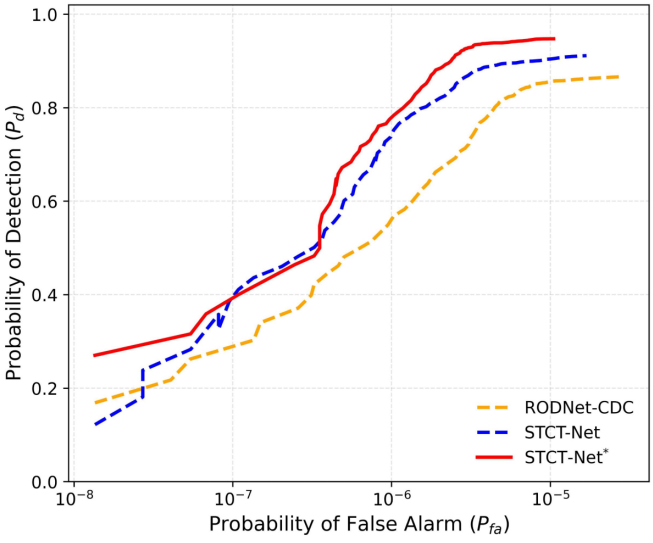

Figure C5 Comparison of the proposed STCT-Net and STCT-Net* with the baseline model RODNet-CDC in terms of ROC curves, with P_d plotted against P_{fa}.

thereby achieving a better balance between detection accuracy and recall performance.

Appendix C.4 Discussion and visualizations

The preceding quantitative analysis systematically verifies the effectiveness of each key component. However, metric-based comparisons alone are insufficient to fully reveal the mechanisms underlying the observed performance improvements. Therefore, this section further discusses the operating principles, scenario adaptability, and potential limitations of the proposed method by combining ablation results with representative visualization cases. First, the feature outputs of different branches within the STCT module are visualized and analyzed. As illustrated in Figure C3, three scenarios with increasing levels of difficulty are presented from top to bottom: longitudinal maneuvering, lateral maneuvering, and hovering. Due to the weak nature of UAV echoes and their susceptibility to clutter and noise interference, the target responses in the original RA features are relatively weak. After temporal branch processing, the model effectively captures motion-related temporal information across consecutive chirps, while the spatial branch further enhances structural representations in the RA plane. Following spatio-temporal fusion, the target responses become more concentrated, while irrelevant background activations are suppressed. Compared with the original input features, the STCT outputs exhibit stronger target discriminative capability and maintain better consistency with the Ground Truth locations. These visualization results demonstrate that the STCT module effectively integrates temporal and spatial information, providing more discriminative feature representations for weak UAV echo detection in the subsequent detection head.

Furthermore, to further evaluate the robustness of the proposed method under challenging detection conditions, we compared the

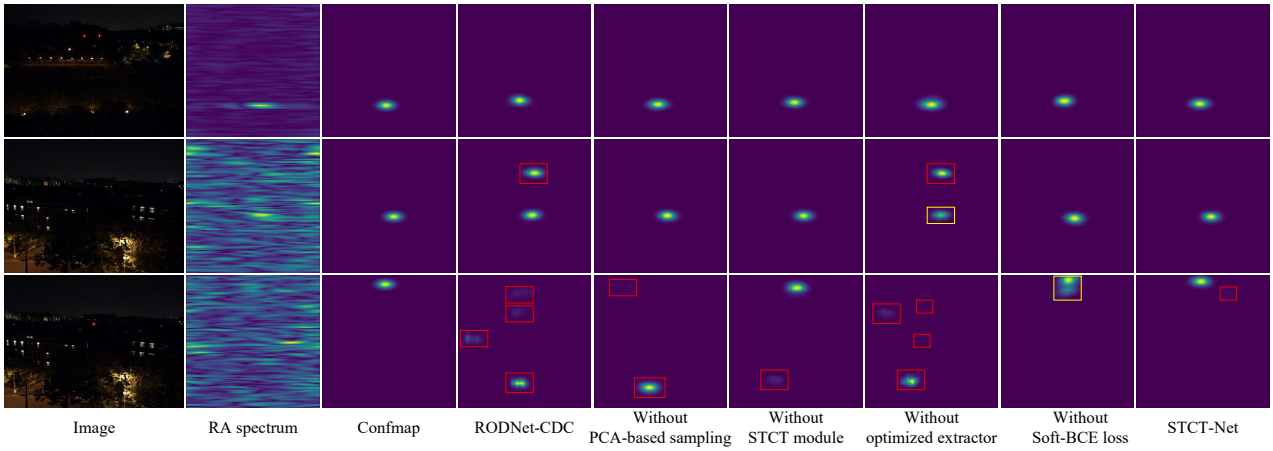

Figure C6 Visualization of the detection results of the baseline model, the ablation models with each key module removed, the complete STCT-Net, and the corresponding ground truth (ConfMap) under high-SNR, low-SNR, and extremely challenging scenarios. Yellow solid boxes indicate ambiguous detections or localization errors, whereas red solid boxes indicate false positives.

performance of RODNet-CDC, STCT-Net, and STCT-Net* across different SNR levels and target ranges, as shown in Figure C4. Overall, as the SNR decreases, target echoes become increasingly obscured by noise and background interference, thereby increasing the difficulty of detection. Compared with RODNet-CDC, STCT-Net and STCT-Net* consistently achieve superior performance under low-SNR conditions, demonstrating their ability to extract discriminative features from weak targets while effectively suppressing noise interference. In particular, STCT-Net* applies PCA to preprocess the chirp data, reducing redundant information and emphasizing dominant echo components, which further enhances weak-target detection under low-SNR conditions. Notably, when the SNR exceeds 20 dB, the detection performance exhibits fluctuations and may even decline. This behavior may be attributed to strong echoes and multipath-induced artifacts in urban environments, which can become more pronounced at high SNRs, increasing target-artifact confusion and false detections. Consequently, a higher SNR does not necessarily translate into improved detection performance. With respect to target range, echo energy generally decreases as the target moves farther from the radar, resulting in increased detection difficulty. However, target range is not necessarily the primary determinant of detection performance. Some short-range targets may also exhibit relatively poor detection performance, primarily because of the UAV flight state. For example, during hovering or lateral motion, UAV radar signatures may become less distinctive and therefore more difficult to distinguish from background interference. Despite the combined effects of SNR, target range, and flight state, STCT-Net and STCT-Net* consistently maintain better detection performance than the baseline model, demonstrating their enhanced capability to represent and detect weak targets in complex scenarios.

In addition, to further investigate the trade-off between target detection and false-alarm suppression, we computed the receiver operating characteristic (ROC) curves of RODNet-CDC, STCT-Net, and STCT-Net* using an OLS threshold of 0.5, consistent with the criterion adopted for mAP_{50}, as shown in Figure C5. As the false-alarm probability (P_{fa}) increases, the detection probability (P_d) of all models increases accordingly. Compared with the baseline RODNet-CDC, STCT-Net and STCT-Net* exhibit consistently higher ROC curves, indicating that the proposed methods achieve higher detection probabilities at the same false-alarm level. In particular, STCT-Net* further outperforms STCT-Net, demonstrating the effectiveness of the corresponding optimization strategy in enhancing detection performance.

Finally, to more intuitively illustrate the role of each module, we further visualized the prediction results of the baseline model, the ablation models obtained by removing key components one at a time, and the complete STCT-Net. The results are shown in Figure C6. The first row presents detection results under a high-SNR scenario. Because the target echoes are relatively clear, all models can localize the UAV with reasonably high accuracy. The second row shows a successful detection case under low-SNR conditions. Compared with the high-SNR scenario, the performance differences among the models become more pronounced. It can be observed that both the baseline model, RODNet-CDC, and the model without the optimized feature extractor exhibit obvious false alarms, indicating that the optimized feature extractor plays an important role in extracting discriminative features for small and weak targets. The third row presents a failure case in an extremely challenging scenario. In this case, all methods exhibit varying degrees of false alarms or localization errors. Yellow solid boxes indicate detections with ambiguity or positional deviation, whereas red solid boxes indicate false alarms. Nevertheless, the complete STCT-Net generally assigns lower confidence scores to incorrect targets, suggesting stronger false-alarm suppression capability. Overall, the performance improvement of STCT-Net results from the synergistic interaction among input representation optimization, spatio-temporal feature modeling, enhanced feature extraction, and improved supervision design, and this advantage is particularly evident in low-SNR and complex clutter scenarios.

References

1 Wang Y, Jiang Z, Li Y, et al. 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, 2021, 15(4): 954–967
2 Fang Y, Shi F, Wei X, et al. CubeDN: Real-time drone detection in 3D space from dual mmWave radar cubes. In: IEEE International Conference on Robotics and Automation, Atlanta, GA, USA, 2025. 3977–3983
3 Zhao P, Lu C X, Wang B, et al. 3D motion capture of an unmodified drone with single-chip millimeter wave radar. In: IEEE International Conference on Robotics and Automation, Xi'an, China, 2021. 5186–5192
4 Chen X, Yuan W, Du X, et al. Multiband FMCW radar LSS-target detection dataset (LSS-FMCWR-1.0) and high-resolution micromotion feature extraction method. Journal of Radars, 2024, 13(3): 539–553
5 Yuan W, Chen X, Du X, et al. A low slow small target classification network model based on K-band radar dynamic multifeature data fusion. IEEE Sensors Journal, 2025, 25(3): 1656–1668

6 Lin Q, Qin Z, Jiang L, et al. Ground-based radar detection dataset of "low slow small" unmanned aerial vehicles under simple field background conditions. Journal of Signal Processing, 2024, 40(11): 2095–2104
7 Zhang L, Tu G, Xu Y, et al. A lightweight CNN-based method for micro-doppler feature-based UAV detection and classification. Electronics. 2025, 14: 4831
8 Yuan S, Yang Y, Nguyen T H, et al. MMAUD: A comprehensive multi-modal anti-UAV dataset for modern miniature drone threats. In: IEEE International Conference on Robotics and Automation, Yokohama,, Japan, 2024. 2745–2751
9 Yu W, Si C, Zhou P, et al. MetaFormer baselines for vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 896–912
10 Liu Z, Mao H, Wu C-Y, et al. A ConvNet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022. 11976–11986
11 Law H and Deng J. CornerNet: Detecting objects as paired keypoints. In: Computer Vision-ECCV 2018: 15th European Conference, Munich, Germany, 2018. 8–14
12 Xu W, Lu P, Zhao Y, et al. E-RODNet: Lightweight approach to object detection by vehicular millimeter-wave radar. IEEE Sensors Journal, 2024, 24(20): 33091–33100
13 Gao X, Xing G, Roy S, et al. RAMP-CNN: A novel neural network for enhanced automotive radar object recognition. IEEE Sensors Journal, 2021, 21(4): 5119–5132
14 Zhuang L, Yao Y, Li N, et al. RC-ROSNet: Fusing 3D Radar Range-Angle Heat Maps and Camera Images for Radar Object Segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2026, 36(1): 421–434
15 Jiang T, Zhuang L, An Q. T-RODNet: Transformer for vehicular millimeter-wave radar object detection. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1–12
16 Dalbah Y, Lahoud J, and Cholakkal H. TransRadar: adaptive-directional transformer for real-time multi-view radar semantic segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2024, 353–362
17 A. Jalil, H. Yousaf and M. I. Baig. Analysis of CFAR techniques. In: 2016 13th International Bhurban Conference on Applied Sciences and Technology, Islamabad, Pakistan, 2016. 654–659
18 P. Addabbo, S. Han, D. Orlando and G. Ricci. Learning Strategies for Radar Clutter Classification. IEEE Transactions on Signal Processing, 2021, 69: 1070–1082
19 Lin T-Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327