Extracting Features from Radar Spectrograms Using Deep Learning for Target Detection

Shunlai Lu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1655300664@qq.com

Jianguo Wang
Research Institute of Artificial Intelligence and Data Science
Xi'an Technological University
Xi'an, China
E-mail: wjg_xit@126.com

***Abstract*—Rooted in Convolutional Neural Networks (CNNs), translation invariance inherently imposes a fundamental constraint on their ability to analyze radar spectrograms, resulting in inadequate feature extraction for distant and small targets. To overcome these limitations, this paper proposes IPRadar_Net. This novel model, built on the Transformer architecture, marks a departure from conventional convolutional and hybrid paradigms. The model exploits the Transformer's lack of translation invariance and it's positional encoding to counteract the performance drop-off with increasing target range. IPRadar_Net adopts a Transformer-based backbone for feature extraction, enhanced with an adaptive positional encoding scheme. This scheme is explicitly designed to model the physical attributes of the radar's range-Doppler-angle dimensions, thus ensuring a more faithful representation of the radar data structure. Experimental results on the public RADDet dataset demonstrate that IPRadar_Net achieves a 14.1% improvement in mean Average Precision (mAP) compared to the baseline RadarResNet model. This performance significantly outperforms existing methods, thereby validating the effectiveness of the proposed model in enhancing radar target detection performance.**

Keywords-Radar Object Detection; Feature Extraction; Vision Transformer; Adaptive Positional Encoding; Radar Spectrogram

I. INTRODUCTION

With the evolution of radar imaging technology toward high resolution, the research paradigm for radar target detection is undergoing a significant transformation. In recent years, academic focus has increasingly shifted toward directly extracting features from radar spectrograms to achieve end-to-end target detection. Compared to traditional frameworks relying on statistical signal processing or sparse point clouds, spectrogram-based deep learning methods demonstrate significant advantages. They not only circumvent the complexity of manually designing Constant False Alarm Rate (CFAR) detectors but also effectively overcome issues of information loss and clustering errors caused by threshold truncation in point cloud data. By employing a data-driven approach, this methodology allows for the mining of richer latent information from raw echoes, thereby significantly improving the accuracy of target spatial localization and category recognition [1].

Deep learning has demonstrated vast potential in this domain. Zhang et al[2]. validated CNNs with Radar ResNet, while Wang et al[3]. proposed RODNet, utilizing deformable 3D convolutions to capture temporal motion features. Jiang et al[4]. further developed T-RODNet, CNN-Transformer hybrid aiming to balance local and global modeling. However, these methods often overlook the unique physical semantics of radar dimensions, failing to fully resolve the misalignment between physical characteristics and network inductive biases.

A core theoretical bottleneck persists in mainstream methods that rely on CNN backbones. CNNs assume "translation invariance," treating features as constant regardless of position. However, radar spectrograms are inherently location-sensitive: according to the radar equation, echo power is inversely proportional to the fourth power of distance ($P_r \propto 1/R^4$). Consequently, identical targets exhibit distinct feature distributions at near versus far ranges. CNNs inherently struggle to adapt to this distance-dependent atte-

https://doi.org/10.58244/ijanmc.263677

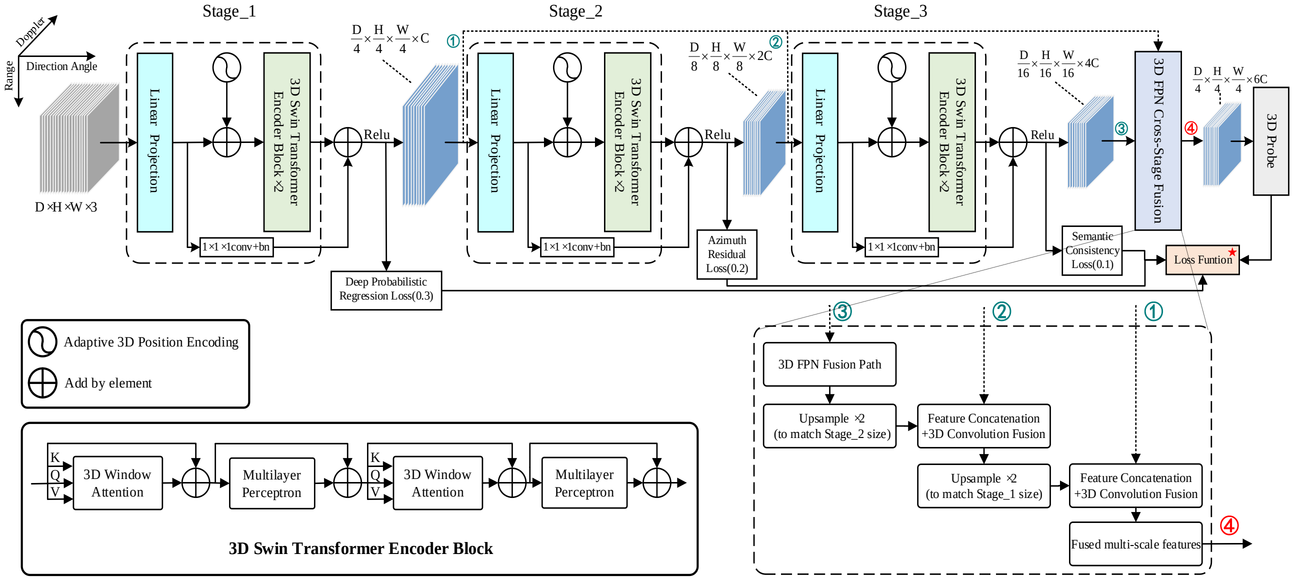
nuation, limiting performance on long-range, weak targets.

To address this, we propose IPRadar_Net, a Transformer-based model. Leveraging the non-translation-invariant nature of Transformers, we discard the convolutional backbone in favor of Adaptive 3D Positional Encoding. This design enables the model to perceive physical semantics across Range, Doppler, and Angle dimensions, dynamically adjusting feature weights. By explicitly preserving spatial physical priors, our approach effectively compensates for echo attenuation, with experiments confirming the superiority of this physics-aware architecture.

II. RADAR TARGET DETECTION MODEL BASED ON BOUNDING BOX DETECTION

A. Network Model Architecture

Given the distinct physical semantics of radar spectrogram dimensions, traditional convolu-tional neural networks exhibit certain limitations in feature representation. To tackle this challenge, this chapter proposes a radar target detection model built upon the Transformer architecture. The block diagram of the model is shown in Fig.1. Drawing on the hierarchical design concept of convolutional neural network models, it adopts a layered feature extraction approach. After the preprocessed RAD image is input into the network, each stage divides the features from the preceding layer into non-overlapping sub-cubes. These are then passed through a positional encoding module and fed into a Swin Transformer encoder for feature trans-formation. Finally, based on the feature map obtained from the last stage, a detection head network is constructed to jointly predict target locations and their corresponding categories.



Stage_1
Stage_2
Stage_3
Doppler
Range
Direction Angle
D ×H ×W ×3
Linear Projection
3D Swin Transformer Encoder Block ×2
1×1 ×1conv+bn
Relu
$\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times C$
①
$\frac{D}{8}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
②
$\frac{D}{16}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
③
3D FPN Cross-Stage Fusion
④
$\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times 6C$
3D Probe
Deep Probabilistic Regression Loss(0.3)
Azimuth Residual Loss(0.2)
Semantic Consistency Loss(0.1)
Loss Funtion
Adaptive 3D Position Encoding
Add by element
K
Q
V
3D Window Attention
Multilayer Perceptron
3D Swin Transformer Encoder Block
3D FPN Fusion Path
Upsample ×2 (to match Stage_2 size)
Feature Concatenation +3D Convolution Fusion
Upsample ×2 (to match Stage_1 size)
Feature Concatenation +3D Convolution Fusion
Fused multi-scale features

Figure 1. The Improved Bounding Box-Based Radar Target Detection Model

In the design of the encoder, the model incorporates an attention residual connection module within the Swin Transformer architecture. This design not only optimizes the gradient propagation path but also significantly enhances the stability of deep network training.

Concurrently, a comprehensive multi-scale training strategy is adopted, leveraging random scale transformation operations to improve the model's feature adaptation capability for diverse multi-scale targets. In the feature fusion module, the constructed 3D feature pyramid network enables dense cross-level feature interaction, effectively fusing rich contextual information across multiple scales. Furthermore, a multi-perspective deep Bayesian fusion mechanism is incorporated into the detection head, which enables the collaborative optimization of multi-source data via probabilistic modeling—thereby significantly enhancing both the confidence of detection results and the overall robustness of the system.

B. Network Architecture Construction

1) Vision Transformer (ViT) Architecture

The Transformer architecture has demonstrated exceptional capabilities in feature extraction, leveraging its self-attention and multi-head attention mechanisms. Subsequent to the achieve-ment of a breakthrough in natural language process-ing, this architecture rapidly expanded into cross-disciplinary fields such as computer vision. Vision Transformer, as a milestone work, success-fully applied a pure Transformer architec-ture to image classification tasks for the first time. To meet the Transformer's requirement for sequential input, ViT divides the input image into regular non-overlapping patches. These patches are flattened and linearly projected into a sequence of tokens via a linear projection layer[5]. To preserve spatial structural information, the model incorporates learnable positional encodings, which are combined with the image tokens through addition. Notably, the model employs a pre-normalization strategy before the attention module. This design not only enhances training stability but also optimizes the adaptation to image feature distributions. Furthermore, this configuration facilitates smoother gradient propagation, thereby accelerating the overall convergence process during the training phase. The detailed architecture is shown in Fig. 2.

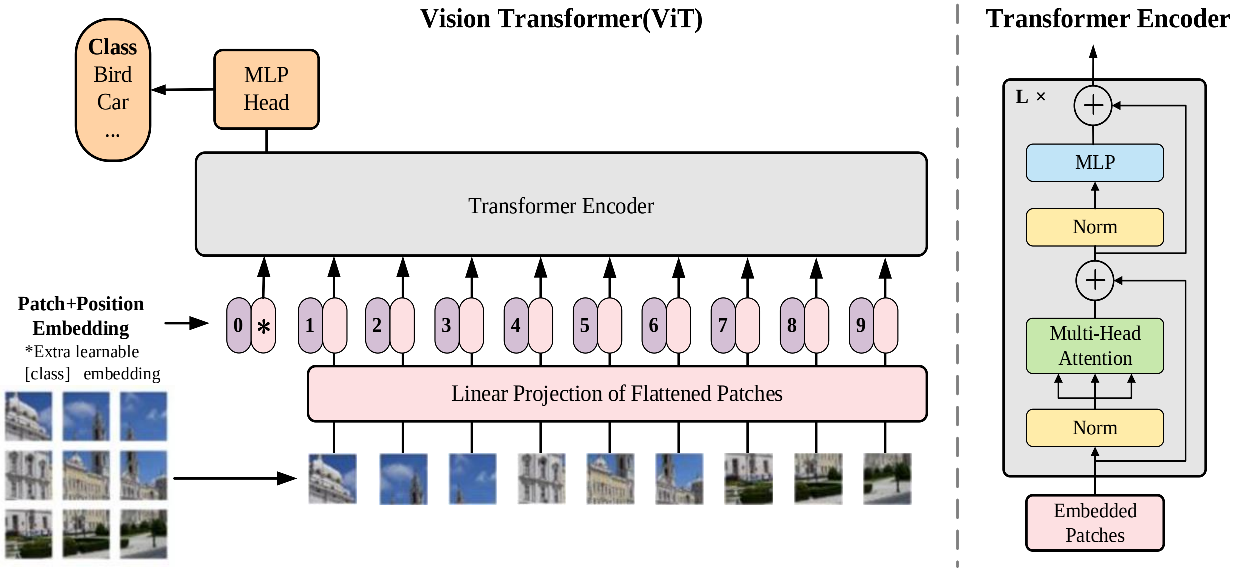


Figure 2. Architecture of the Vision Transformer (ViT) model.

While Vision Transformer has demonstrated superior performance over traditional convolu-tional networks in image classification tasks, its computational complexity increases quadratically with input size, severely limiting its applicability in practical scenarios. The data processed in this study is in a three-dimensional cube format, which would lead to an even more pronounced computational efficiency bottleneck if the standard ViT architecture were directly adopted[6]. To address this, Swin Transformer introduces local window-based self-attention computation, decomposing global atten-tion into operations within fixed-size windows. This design effectively achieves linear computa-tional complexity.

To mitigate the issue of inter-window information isolation that may arise from the local window mechanism, Swin Transformer further proposes a shifted window attention scheme. This mechanism establishes cross-window feature interaction paths by alternating between fixed and shifted window configurations across different network layers.

Specifically, the shifted window operation displaces the window boundaries by half a window size along the spatial dimensions, enabling attention relationships between elements originally in separate windows (as illustrated in the architecture diagram, Fig.3). This hierarchical attention computation pattern significantly enhances the model's ability to capture long-range dependencies while maintaining linear computational complexity.

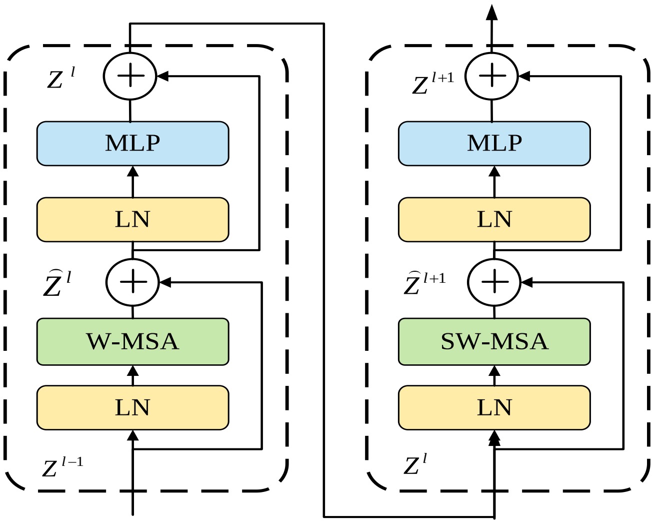


Figure 3. Architecture of the Swin Transformer encoder.

In this model, each hierarchical level re-divides the input features into non-overlapping sequences of 3D patches. The specific partitioning strategy is as follows: the first level employs a patch size of 4×4×4, the second level reduces it to 2×2×2, and by the third level, to accommodate the reduced feature map size along the Doppler dimension, the patch size is further adjusted to 1×2×2. At each level, two 3D Swin Transformer encoder modules are deployed sequentially: the first module performs local window-based self-attention computation, while the second module employs a shifted window attention mechanism that enables cross-window connectivity. Both modules compute self-attention based on 4×4×4 3D windows, with the latter applying a 2×2×2 cyclic shift to the windows prior to computation to facilitate cross-window information exchange.

2) *Multi-Stage 3D Swin Transformer Encoder*

The multi-stage encoder employs a physics-aware asymmetric partition strategy to accommodate radar characteristics. Since the Doppler dimension possesses low resolution yet encodes critical micro-motion features, isotropic down-sampling risks losing velocity information. Therefore, we utilize patch sizes 4×4×4 and 2×2×2 in shallow layers, adjusting to 1×2×2 in deep layers. This preservation of the Doppler dimension prevents the loss of physical details, significantly enhancing the discrimination of slow-moving and micro-motion targets from complex environmental background clutter.

Subsequently, we propose an Adaptive 3D Positional Encoding mechanism to explicitly model spatial relationships. Unlike natural images, radar spectrogram dimensions (Range, Doppler, Angle) carry distinct physical semantics where spatial distribution correlates with systematic feature patterns[7]. This method dynamically adjusts the importance of each dimension based on input features[8].

Let the input feature be X. The adaptive positional encoding is expressed as:

$$PE = \alpha(X)pos_h + \beta(X)pos_w + \eta(X)pos_d \quad (1)$$

Where the positional encodings along the three dimensions are denoted as pos_h、pos_w、pos_d , separately. Each dimensional encoding employs sine and cosine functions with different frequencies, given by equation (2) and (3):

$$pos(pst,2i) = sin\left(\frac{pst}{10000^{\frac{2i}{D}}}\right) \quad (2)$$

$$pos(pst,2i+1) = cos\left(\frac{pst}{10000^{\frac{2i}{D}}}\right) \quad (3)$$

Where *pos* denotes the position of the current block within the entire sequence, and *D* represents the dimension of the sequence[9]. α, β and γ are the adaptive scale factors for the three dimensions, which are dynamically generated by a specifically designed gating function $g(X)$. The formal definition of the linear weight calculation process is presented in Equation (4):

$$[\alpha,\beta,\gamma]=g(X)=\sigma(W_2\cdot\delta(W_1\cdot\mathrm{GAP}(X))) \quad (4)$$

$\mathrm{GAP}(\cdot)$ denotes Global Average Pooling, utilized to compress spatiotemporal information; W_1 and W_2 represent learnable weight matrices; and δ corresponds to the ReLU activation function. The Sigmoid activation function is selected at the output because it strictly maps values to the (0,1) interval, endowing the scale factors $\alpha,\beta,$ and γ with a clear physical interpretation of "gating probability." This enables the model to adaptively learn the importance of each physical dimension in a data-driven manner, thereby controlling the relative contribution of positional information from each axis to the feature representation.

3) 3D FPN Cross-Scale Feature Fusion Module

Based on the features output by the multi-stage encoder, which belong to different scales (fine-grained, medium-grained, and coarse-grained), a single-scale feature struggles to simultaneously meet the detail requirements for small object detection and the global context demands for large object detection. Therefore, a 3D Feature Pyramid Network (FPN) cross-stage fusion module is designed to achieve effective multi-scale feature integration[10].

The 3D FPN employs a three-level architecture comprising top-down upsampling, lateral feature alignment, and nonlinear fusion, with its operators optimized for 3D spatial properties to maintain the integrity of spatial and semantic information.

During the process of feature representation and lateral alignment, let the output features from each stage of the multi-stage encoder be denoted $F_i\in R^{D_i\times H_i\times W_i\times C_i}\ (i=1,2,3)$, where $D_i=D/4^i$, $H_i=H/4^i$, and $W_i=W/4^i$ represent the three spatial dimensions, and $C_i=C\times 2^{i-1}$ is the number of feature channels. Owing to discrepancies in both the channel count and spatial dimensions across different stages, lateral connections are employed to achieve dimension unification and feature alignment.

This is accomplished through a $1\times1\times1$ 3D convolution that performs channel mapping on the features from each stage, which standardizes the channel dimensions by unifying features with different channel counts into a common fusion channel dimension $C_{fusion}=2C$. The mapping is defined by Equation (5):

$$F_i^{'}=Conv_{1\times1\times1}(F_i;\theta_i)=\sigma(W_i*F_i+b_i) \quad (5)$$

Where $\theta_i=\{W_i,b_i\}$ represents the learnable parameters of the i-th lateral convolutional layer (with kernel $W_i\in R^{1\times1\times1\times C_i\times C_{fusion}}$ and bias $b_i\in R^{C_{fusion}}$), * denotes the 3D convolution operation, σ is the activation function, and $F_i^{'}\in R^{D_i\times H_i\times W_i\times C_{fusion}}$ denotes the channel-aligned features.

This is followed by upsampling operations to achieve spatial size matching between deep and shallow features, ensuring spatial dimensional consistency and thereby laying a solid foundation for element-wise fusion.

Throughout the top-down progressive fusion process, starting from the deepest features (Stage_3), upsampling is performed via 3D transposed convolution. The upsampled features are then progressively integrated with those from earlier stages[11]. The specific steps and their mathematical formulations are as follows:

a) Primary Fusion (Stage_3 → Stage_2)

This step employs a $3\times3\times3$ 3D transposed convolution (with s=2) on the channel-aligned features $F_3^{'}$ from Stage_3, achieving a $2\times$ increase in spatial resolution to match Stage_2. The mapping is defined by Equation (6):

$$F_3^{up}=TransConv_{3\times3\times3}(F_3^{'};\phi_3)=W_3^{T}*F_3^{'}+b_3^{T} \quad (6)$$

Where $\phi_3=\{W_3^{T},b_3^{T}\}$ represents the learnable parameters of the transposed convolution layer (with $W_3^{T}\in R^{3\times3\times3\times C_{fusion}\times C_{fusion}}$ being the transposed convolution kernel), and $F_3^{up}\in R^{D_2\times H_2\times W_2\times C_{fusion}}$ denotes the upsampled feature map.

Element-wise addition is performed between feature F_3^{up} and the aligned Stage_2 feature $F_2^{'}$. The sum is then refined by a $3\times3\times3$ 3D convolution for non-linearity and denoising, generating the first-level fused feature, this process is defined by Equation (7):

$$F_{2,\ fusion} = Conv_{3\times3\times3}(F_i^{up} \oplus F_2^{'}; \omega_2) \qquad (7)$$

Where $\omega_3 = \{W_2^{\ f}, b_2^{\ f}\}$ denotes the parameters of the fusion convolutional layer. $\oplus$ signifies the element-wise addition. $F_{2,\ fusion} \in R^{D_2\times H_2\times W_2\times C_{fusion}}$.

b) Secondary Fusion (Stage_2 → Stage_1)

A recurrent application of the transposed convolution upsampl to feature $F_{2,\ fusion}$produces a Stage_1-sized feature map (Equation (8)):

$$F^{up}_{2,\ fusion} = TransConv_{3\times3\times3}(F_{2,\ fusion}; \varphi_2) \ (8)$$

Element-wise fusion of this map with the aligned feature $F_1^{'}$, followed by optimization with a $3\times3\times3$ 3D convolution, ultimately yields the final multi-scale fused feature (Equation (9)):

$$F_{total} = Conv_{3\times3\times3}(F^{up}_{2,fusion} \oplus F_1^{'}; \omega_1) \qquad (9)$$

The fused feature $F_{total} \in R^{D_1\times H_1\times W_1\times C_{fusion}}$preserves Stage_1's high-resolution detail while incorporating Stage_2's structural information and Stage_3's contextual semantics.

Cross-scale fusion achieves optimal feature weighting via linear combination and non-linear transformation[12]. The final feature is formulated as a composite function of the original features (Equation (10)):

$$F_{total} = \tau_1(\mu_2(\tau_2(\mu_3(F_3) \oplus A_2(F_2))) \oplus A_1(F_1)) \qquad (10)$$

Where $A_i(\cdot) = Conv_{1\times1\times1}(\cdot;\ \theta_i)$ denotes the lateral alignment, $\mu_i(\cdot) = TransConv_{3\times3\times3}(\cdot;\ \phi_i)$ represents the upsampling, $\tau_i(\cdot) = Conv_{3\times3\times3}(\cdot;\ \omega_i)$ signifies the non-linear transformation applied after fusion. This formulation captures the hierarchical fusion workflow: upsampled deep features propagate semantics, shallow features supply spatial details, and non-linear transformations model complex cross-scale interactions—producing a highly expressive representation.

4) Bounding Box Prediction

The bounding box detection head employed in this work is based on the anchor mechanism, with a design philosophy consistent with the YOLO series of methods[13]. Specifically, six 3D anchor boxes are obtained by performing K-means clustering on the ground-truth bounding boxes. The detection head itself is constructed by stacking three convolutional layers, serving as a 3D extension of the YOLO detection head.

The final layer has an output channel size of $N \times (box + conf + cls)$ (where NN is the number of anchor boxes; box corresponds to the parameters of the 3D bounding box, including the center point's 3D coordinates $[x, y, z]$ and the box's length, width, and height $[w, h, d]$; conf is the confidence score; and cls is the number of object categories)[14]. Subsequently, the output from the detection head is fed into a Non-Maximum Suppression (NMS) algorithm to eliminate redundant detections, yielding the final detection results.

The prediction tasks of the proposed detection head comprise three components: bounding box regression, confidence prediction, and object classification[13]. Accordingly, the overall loss function consists of three parts, as formulated in Equation (11):

$$L_{total} = \lambda_{box} L_{box} + \lambda_{conf} L_{conf} + \lambda_{cls} L_{cls} \qquad (11)$$

Where:

- λ_{box} denotes the bounding box regression loss, computed using the Generalized Intersection over Union (GIoU) loss. This metric effectively handles non-overlapping bounding boxes and enhances regression accuracy.

- λ_{conf} denotes the confidence loss, implemented with Focal Loss. By reducing the weight of easy-to-classify examples and focusing on hard negatives, it mitigates the issue of foreground-background class imbalance.
- λ_{cls} denotes the classification loss, calculateed via cross-entropy loss, which minimizes the discrepancy between predicted class probabilities and ground-truth labels during the iterative training process.

In the experimental setup, the focusing parameter λ_{box} and the weighting factor α in Focal Loss are set to 0.2 and 0.01, respectively, while the balance coefficients λ_{conf} and λ_{cls} are all set to 1. This combination of parameters is empirically validated to balance the training performance across all detection tasks.

III. STYLING EXPERIMENT AND ANALYSIS

The experiments were implemented using the PyTorch 1.10.0 deep learning framework on a Windows 11 operating system. The hardware computing platform was equipped with an Intel Core i7-9900 CPU @ 3.60 GHz and 32 GB of RAM. Model training and inference were accelerated by an NVIDIA GeForce RTX 3080Ti GPU (12 GB VRAM), configured with the CUDA 11.3 toolkit and the corresponding cuDNN 8.2 acceleration library.

A. Dataset

The RADDet dataset was acquired using a Texas Instruments AWR1843-BOOST millimeter-wave radar sensor paired with stereo cameras. While the radar captured the primary signal data for training, the cameras served to facilitate the ground truth annotation of target positions within the radar spectrograms. The data utilized in this study were collected in dynamic road scenarios. The radar sensor was deployed on the sidewalk with its main beam directed towards the road, covering an effective detection range of 5 to 50 meters. After quality control screening and the removal of invalid frames, a total of 10,158 valid frames were retained. The dataset was partitioned into training and testing sets at a ratio of 8:2. Each data frame comprises three types of information: raw ADC sampling signals, processed RAD spectrogram data, and synchronized optical images. For model input, one can either directly utilize the RAD spectrograms or preprocess the raw ADC signals using standard signal processing techniques to generate the required feature representations.

The dataset includes six target categories: Person, Car, Bicycle, Motorcycle, Truck, and Bus. The class distribution exhibits significant imbalance: Person and Car samples dominate the dataset, whereas Motorcycle and Bus samples are scarce. Such distributional imbalance is anticipated to adversely affect the detection performance for minority classes. Annotations are provided in bounding box format, offering two distinct systems: 3D bounding boxes defined within the RAD spectrograms, and 2D bounding boxes generated in the bird's-eye view (BEV) after transforming the Range-Azimuth map into the Cartesian coordinate system. Both annotation types precisely provide the target's center coordinates, dimensions, and semantic category. Given that the core task of this study is target localization within RAD spectrograms, only the 3D bounding box annotations are adopted as the ground truth for model training and evaluation.

B. Implementation Details

For the bounding box detection task on the RADDet dataset, the model was trained for a total of 100 epochs. The AdamW optimizer was employed to decouple weight decay from gradient updates, with the weight decay coefficient set to 1×10^{-4} and the batch size set to 4. The initial learning rate was set to 0.0005. To ensure convergence stability in the later stages of training, a multi-step learning rate decay strategy was adopted: the learning rate was decayed to 0.1 times its current value (decay factor $\gamma = 0.1$) at the 35th and 50th epochs, respectively. During the testing phase, the confidence threshold was set to 0.5, and the Non-Maximum Suppression (NMS) threshold was set to 0.1.

C. Data Preprocessing

In the data preprocessing phase, raw radar ADC signals must be converted into Range-Azimuth-Doppler (RAD) spectrograms suitable for model input. This study utilizes the RADDet dataset, where the raw data possesses dimensions of $64\times256\times8$. The preprocessing pipeline initiates by sequentially performing Fast Fourier Transforms (FFT) along the range, Doppler, and azimuth dimensions, completing the time-frequency domain transformation to generate the RAD spectrogram.

Given the complex nature of the resulting data, it is separated into real and imaginary components as independent channels. Additionally, the base-10 logarithm of the amplitude is computed. These components are combined to form a 3D data structure with dimensions of $64\times256\times256\times3$. Finally, the data undergoes Z-Score normalization to mitigate the impact of numerical distribution discrepancies across different dimensions on gradient propagation. The complete expression for the revised normalization formula is presented in Equation (12):

$$\hat{x}=\frac{x-\mu}{\sigma+\epsilon} \tag{12}$$

Where x represents the raw spectrogram features after logarithmic transformation, while μ and σ denote the channel mean and standard deviation calculated across the entire training set, respectively. ϵ is a small constant (set to 1×10^{-5} in this study) introduced to prevent division by zero. Following normalization, the input is adjusted to a standard normal distribution with zero mean and unit variance, significantly enhancing the convergence speed and stability.

D. Performance EvaluationData

1) Enhanced Model Performance Evaluation

This section evaluates bounding box detection performance on the RADDet dataset using RAD spectrograms. RadarResNet served as the baseline model under our experimental setup, while this paper proposed detection framework, designated IPRadar_Net, was trained for 100 epochs (Fig.4)

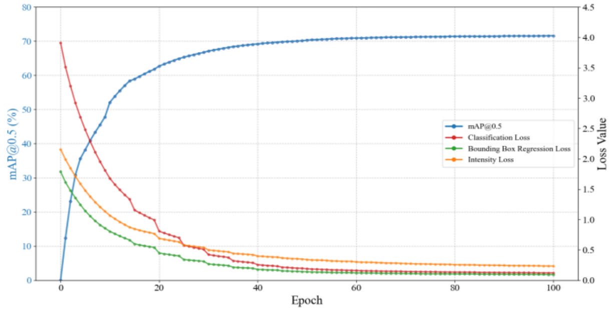


Figure 4. Evolution of mAP@0.5 and Multi-task Loss Functions

During evaluation, four Intersection over Union (IoU) thresholds (0.1, 0.3, 0.5, 0.7) were adopted to calculate category-wise Average Precision (AP) and mean Average Precision (mAP), allowing for a comprehensive assessment of the model's localization performance across varying strictness levels. All reported results are based on the model achieving peak mAP on the validation set during the complete 100-epoch training cycle, to ensure the selection of the best-performing checkpoint as evaluated on the test set. The detailed quantitative results corresponding to these evaluations are comprehensively summarized in Table I and Table II.

TABLE I. EXPERIMENTAL RESULTS OF THE RADDET DATASET

RadarResNet	$AP^{0.1}$	$AP^{0.3}$	AP$^{0.5}$	$AP^{0.7}$
person	0.712	0.344	0.045	0.002
car	0.856	0.683	0.351	0.069
bicycle	0.531	0.343	0.043	0
bus	0.602	0.477	0.238	0
truck	0.667	0.572	0.286	0.05
motorcycle	0.144	0.144	0.093	0
mAP	0.585	0.427	0.176	0.02

TABLE II. EXPERIMENTAL RESULTS OF IPRADAR_NET DATASET

IPRadar_Net	$AP^{0.1}$	$AP^{0.3}$	AP$^{0.5}$	$AP^{0.7}$
person	0.843	0.454	0.121	0.002
car	0.956	0.783	0.449	0.118
bicycle	0.764	0.53	0.125	0
bus	0.813	0.672	0.438	0.186
truck	0.867	0.772	0.486	0.127
motorcycle	0.474	0.442	0.193	0.083
mAP	0.786	0.609	0.302	0.086

IPRadar_Net surpasses RadarResNet with mAP gains of 0.201/0.182/0.126/0.066 at IoU = 0.1/0.3 /0.5/0.7(mean+0.144).

Comprehensive AP-IoU curves for all six categories on the RADDet dataset demonstrate IPRadar_Net's consistent superiority over the RadarResNet baseline. The proposed model (solid lines) significantly outperforms the baseline (dashed lines) across the entire IoU spectrum [0.10-0.95] in all object categories, as visually confirmed by the comparative plotting with color-coded category matching.

Fig.5 shows complete AP-IoU curves for all six categories on RADDet, with color-coded categorie comparing IPRadar_Net (solid) versus RadarResNet baseline (dashed). IPRadar_Net achieves significant performance gains across all IoU thresholds, demonstrating particular strength on radar-challenging small targets: 2-4×AP improvement at IoU=0.5 for pedestrians (blue), bicycles (green) and motorcycles (pink), with sustained superiority in medium-high IoU ranges (0.3-0.7).

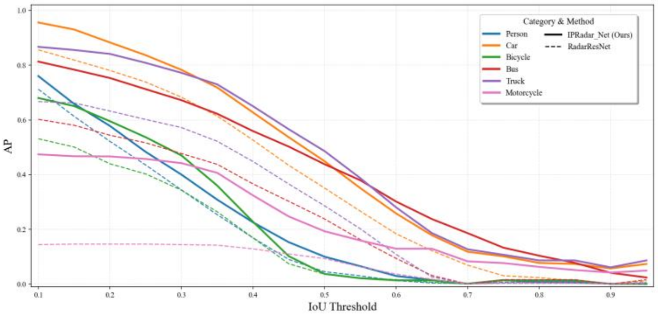


Figure 5. AP–IoU Curves of IPRadar_Net and RadarResNet on the RADDet Dataset.

These results confirm IPRadar_Net's enhanced sparse feature modeling capability and robust small-object detection in complex traffic scenarios. Evaluation follows official RADDet protocol and mainstream radar/LiDAR detection standards.

TABLE III. SUMMARY OF 4-POINT MAP METRICS FOR THE IPRADAR_NET DATASET

Category	*mAP*		*Enhance*
	RadarResNet	*IPRadar_Net*	
Person	0.276	0.368	+0.092
Car	0.490	0.672	+0.182
Bicycle	0.229	0.355	+0.126
Bus	0.329	0.532	+0.203
Truck	0.383	0.581	+0.198
Motorcycle	0.095	0.298	+0.203

Based on Table III, Fig.6 presents a 4-point mAP radar chart comparison on RADDet using IoU thresholds [0.1, 0.3, 0.5, 0.7]. The proposed IPRadar_Net (solid red) achieves 2.1×larger radar area than the RadarResNet baseline (purple), demonstrating significant perfor-mance gains across all categories. Particularly for small targets such as persons, bicycles, and motorcycles, IPRadar_Net achieves a significant performance gain, improving the mAP by a factor of 2 to 3. This substantial increase validates the comprehensive superiority of the model in multi-precision object detection.

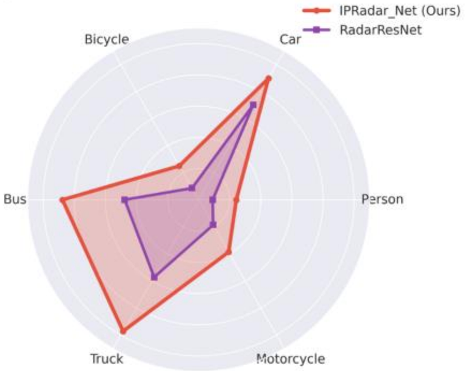
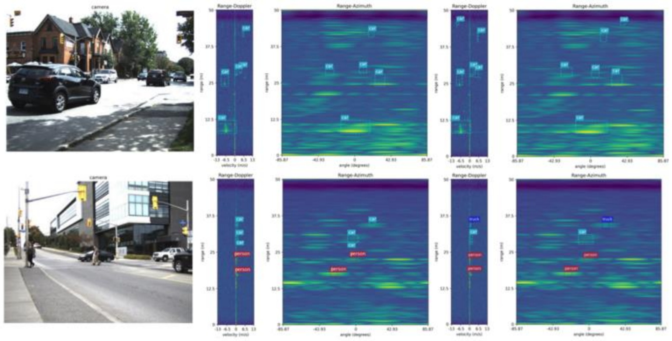


Figure 6. DatasetPer-class AP@0.5 Radar Chart(RADDet)

All categories show consis-tent AP improvements.



(a) Original image (b)Radar ResNeet (c)IPRadar_Net

Figure 7. Visualization of the RADDet Dataset

Visual results (Fig.7) confirm superior detection accuracy, with reduced missed/ false detections and more precise localization, demonstrating enhanced feature representation for radar data.

To further validate the effectiveness of the proposed model, this experiment applies the ViT model to bounding box detection and central keypoint detection tasks, and compares its performance with the model introduced in this chapter.

TABLE IV. PERFORMANCE COMPARISON BETWEEN VIT AND IPRADAR_NET ON RADDET DATASET

Model	$mAP^{0.1}$	$mAP^{0.3}$	$mAP^{0.5}$	$mAP^{0.7}$
ViT	0.647	0.493	0.225	0.032
IPRadar_Net	0.786	0.584	0.294	0.053

As validated in Table IV, the proposed architecture consistently surpasses the Vision Transformer (ViT) across all metrics, revealing two fundamental limitations of ViT in radar processing: its global self-attention mechanism expends excessive computation on spectrogram-wide processing despite targets being local extrema, and its inherent lack of multi-scale feature extraction constrains representational capacity. These structural disadvantages substantiate the performance gap while justifying our architectural modifications.

2) *Impact of Adaptive 3D Position Encoding*

The proposed framework incorporates adaptive 3D positional encoding that independently processes three spatial dimensions with dynamically scaled factors during training.

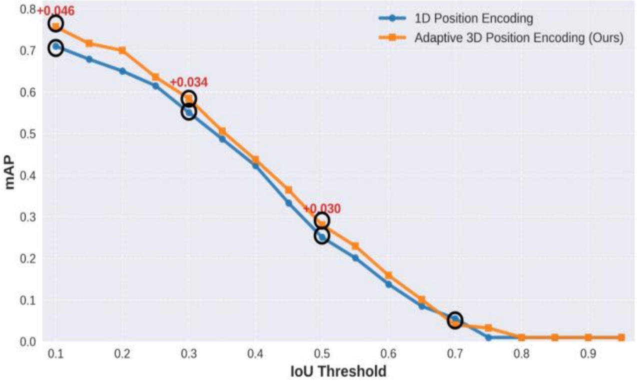


Figure 8. Ablation Study：Adaptive 3D VS 1D Position Encoding

Ablation studies replacing this 3D encoding with standard 1D encoding-under identical experimental conditions on RADDetde-monstrate con-sistent performance improvements across multiple IoU thresholds, as quantitatively verified in Fig.8.

Ablation studies confirm the critical advantage of our adaptive 3D encoding over standard 1D encoding. As quantified in Table V, substituting 1D encoding with our 3D approach yields mAP improvements of 0.059, 0.031, and 0.036 at IoU thresholds of 0.1, 0.3, and 0.5 respectively on the RADDet benchmark. These consistent gains validate that dimension-aware encoding provides significantly more meaningful positional cues to the Transformer architecture, consequently enhancing detection accuracy. The substantial performance improvements demonstrably justify the modest additional computational overhead introduced during the entire training process.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT POSITION ENCODING METHODS ON RADDET DATASET

Positional Encoding	$mAP^{0.1}$	$mAP^{0.3}$	$mAP^{0.5}$	$mAP^{0.7}$
1D Position Encoding	0.706	0.553	0.255	0.051
Adaptive 3D Position Encoding	0.765	0.584	0.291	0.051

3) *Impact on Category Overlap in RADDet Dataset*

Evaluation on the RADDet dataset indicates that the motorcycle class significantly compromises overall detection performance due to its limited samples, with the bus category similarly affected by data scarcity. Given that standard radar detection benchmarks typically contain only three categories (pedestrian, cyclist, and car), we conducted a category consolidation experiment: motorcycles were merged into cyclists, while buses and trucks were merged with cars. The proposed IPRadar_Net was then re-evaluated under this new taxonomy, with the specific experimental results presented in Table VI.

TABLE VI. PERFORMANCE COMPARISON OF DIFFERENT POSITION ENCODING METHODS ON RADDET DATASET

IPRadar_Net	$AP^{0.1}$	$AP^{0.3}$	$AP^{0.5}$	$AP^{0.7}$
person	0.736	0.432	0.132	0.012
car	0.928	0.782	0.327	0.082
truck	0.867	0.715	0.457	0.097
mAP	0.844	0.643	0.305	0.064

Category merging improved only mAP@0.1 (due to motorcycle removal) but hurt other metrics. Merged classes suffered from high intra-class variance in RCS/size, reducing detection accuracy. Consolidation proves ineffective for RADDet.

4) Comparative Analysis of Comprehensive Performance and Efficiency

To comprehensively evaluate the engineering practicality and advancement of IPRadar_Net, this section conducts a comparative analysis of the comprehensive performance of various models from the dual dimensions of computational efficiency and detection accuracy. In addition to the baseline RadarResNet and the pure Transformer architecture (ViT), the mainstream CNN-Transformer hybrid architecture, T-RODNet, is explicitly introduced as a critical benchmark. This comparison aims to provide an in-depth investigation into the trade-offs between resource consumption and perception performance across different architectures.

Table VII demonstrates that IPRadar_Net achieves an optimal trade-off between computational efficiency and detection accuracy.

TABLE VII. COMPREHENSIVE PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE RADDET DATASET

Model	*Backbone*	*Params/ M*	*FLOPs/ G*	*FPS*	*mAP/ %*
RadarResNet	CNN	2.5	15.2	45.2	58.5
ViT	ViT-B	86.4	120.5	12.1	64.7
T-RODNet	CNN + Transformer	44.3	182.5	14.5	72.4
IPRadar_Net	Swin-T + Adaptive PE	28.3	45.6	28.5	78.6

Although the inference speed (28.5 FPS) is slightly lower than that of pure CNN architectures due to the self-attention mechanism, it significantly surpasses the physical refresh rate of mainstream radars (typically 10－20 Hz), thereby fully satisfying real-time requirements. Furthermore, benefiting from the linear complexity of the Swin Transformer, the inference efficiency is improved by approximately 2.4 times compared to the standard ViT. In comparison with the hybrid T-RODNet, our model — by discarding the computationally intensive 3D CNN backbone in favor of Adaptive 3D Positional Encoding — achieves a 6.2% increase in mAP and nearly doubles the inference speed, while reducing the parameter count by 36%. This result validates that direct physical spatial modeling using Transformers is superior to the traditional "CNN-Spatial + Transformer-Temporal" hybrid paradigm in effectively capturing the weak features of distant targets.

IV. CONCLOUSIONS

This study systematically investigates radar target detection, emphasizing the distinct physical meanings in radar spectrogram dimensions and feature variations due to target spatial differences. CNN-based models, limited by translation invariance, struggle to represent radar features effectively, hindering performance improvements. We introduce IPRadar_Net, a Transformer-based bounding box detector, as an enhanced solution for radar target detection.

The model advances beyond traditional limits with two key improvements. It uses Transformer for feature extraction, exploiting its lack of translation invariance and positional encoding to better handle radar spectrogram challenges, matching radar data's physics. It also includes an adaptive 3D positional encoding module for precise Transformer positioning and optimizes the center keypoint decoder by fusing self- and cross-attention for stronger feature interactions. These changes boost feature capture and use, outperforming CNNs and hybrids in specificity and advancement.

Experiments on the RADDet dataset validate the approach, with mAP rising 14.1% over the baseline Radar ResNet, gains in multiple metrics, and overall performance exceeding state-of-the-art methods. This confirms the Transformer's superiority in adapting to radar data and overcoming CNN issues, highlighting the encoding module and decoder optimizations' roles in feature enhancement.

Future work involves lightweight designs incorporating sparse attention mechanisms and anti-interference enhancements to improve robustness in complex environments.

REFERENCES

[1] Research on RODNet-based Millimeter Wave Radar Target Detection Using Gaussian Mixture Model [D]. Li Zhuang. North China University of Technology, 2023.

[2] Zhang W, Li H, Sun G, et al. Enhanced detection of Doppler-spread targets for FMCW radar[J]. IEEE Transactions on Aerospace and Electronic Systems, 2019, 55(4): 2066-2078.

[3] Wang Y, Wang G, Hsu H M, et al. Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2815-2824.

[4] Jiang T, Zhuang L, An Q, et al. T-rodnet: Transformer for vehicular millimeter-wave radar object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 72: 1-12.

[5] GHASR M T, KHARKOVSKY S, BOHNERT R, et al. 30 GHz Linear High-Resolution and Rapid Millimeter Wave Imaging System for NDE[J]. IEEE Transactions on Antennas and Propagation, 2013,61(9): 4733-4740.

[6] Review of Automatic Detection and CFAR Processing Method of Radar [J]. He You, Key, Meng Xiangwei, Lu Da, Peng Yingning. System Engineering and Electronic Technology, 2001 01)

[7] KHARKOVSKY S, CASE J T, ABOU-KHOUSA M A, et al. Millimeter-Wave Detection of Localized Anomalies in the Space Shuttle External Fuel Tank Insulating Foam[J]. IEEE Transactions on Instrumentation and Measurement, 2006,55(4): 1250-1257.

[8] PHAM T, KIM K, HONG I. A Study on Millimeter Wave SAR Imaging for Non-Destructive Testing of Rebar in Reinforced Concrete[J]. Sensors, 2022,22(20): 8030.

[9] GAO Y, ZOUGHI R. Millimeter Wave Reflectometry and Imaging for Noninvasive Diagnosis of Skin Burn Injuries[J]. IEEE Transactions on Instrumentation and Measurement, 2017,66(1): 77-84.

[10] Cheng Yufeng. Research on Traffic Adaptability of Roadside Sensors in Vehicle-to-Infrastructure (V2I) Environment [D]. Beijing: Beijing Jiaotong University, 2021.

[11] Lauteslager T, Tømmer M, Lande T S, et al. Dynamic microwave imaging of the cardiovascular system using ultra-wideband radar-on-chip devices[J]. IEEE Transactions on Biomedical Engineering, 2022, 69(9): 2935-2946.

[12] Cheng Y, Liu Y. Person reidentification based on automotive radar point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-13.

[13] Meng Z, Fu S, Yan J, et al. Gait recognition for co-existing multiple people using millimeter wave sensing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 849 856.

[14] Zhang A, Nowruzi F E, Laganiere R. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users[C]//2021 18th Conference on Robots and Vision (CRV). IEEE, 2021: 95-102