End-to-End Radar Human Segmentation with Differentiable Positional Encoding

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

TR2025-125 August 26, 2025

Abstract

The Radar dEtection TRansformer (RETR) has recently been introduced to fuse multi-view millimeter-wave radar heatmaps using a detection transformer framework and a simple geometric learning approach for indoor radar perception. A key part of RETR is its tunable positional encoding (TPE), which adjusts the weight of depth positional embeddings across different views to improve feature matching. However, the original design fixes the TPE ratio before training. Differentiable Positional Encoding (DiPE) was proposed to overcome this limitation for bounding box detection by automatically adjusting the TPE ratio with dual differentiable masks on depth and angular positional embeddings. In this paper, we build on the existing DiPE approach and propose a segmentation pipeline that extends its application to human instance segmentation directly from radar signals. Our method integrates the established DiPE mechanism into a framework for segmentation, working with either fixed (e.g., sinusoidal) or learnable positional embeddings, and is optimized end-to-end with a segmentation loss. Evaluation on the open-sourced MMVR dataset shows that our segmentation pipeline achieves improved performance compared to conventional methods.

European Signal Processing Conference (EUSIPCO) 2025

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

End-to-End Radar Human Segmentation with Differentiable Positional Encoding

Ryoma Yataka[1], Pu (Perry) Wang[2], Petros Boufounos[2], and Ryuhei Takahashi[1]

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

***Abstract*—The Radar dEtection TRansformer (RETR) has recently been introduced to fuse multi-view millimeter-wave radar heatmaps using a detection transformer framework and a simple geometric learning approach for indoor radar perception. A key part of RETR is its tunable positional encoding (TPE), which adjusts the weight of depth positional embeddings across different views to improve feature matching. However, the original design fixes the TPE ratio before training. Differentiable Positional Encoding (DiPE) was proposed to overcome this limitation for bounding box detection by automatically adjusting the TPE ratio with dual differentiable masks on depth and angular positional embeddings. In this paper, we build on the existing DiPE approach and propose a segmentation pipeline that extends its application to human instance segmentation directly from radar signals. Our method integrates the established DiPE mechanism into a framework for segmentation, working with either fixed (e.g., sinusoidal) or learnable positional embeddings, and is optimized end-to-end with a segmentation loss. Evaluation on the open-sourced MMVR dataset shows that our segmentation pipeline achieves improved performance compared to conventional methods.**

***Index Terms*—Indoor radar perception, instance segmentation, radar detection transformer, positional encoding.**

I. Introduction

Radar provides robust and reliable detection in low light, adverse weather and hazardous conditions at a lower cost than cameras and LiDAR. Its applications have expanded from outdoor automotive sensing [1]–[5] to indoor scenarios such as elder care, energy management, and navigation [2], [6]–[8]. However, the extraction of fine-grained semantic features from radar signals remains challenging. Early works relied on low-resolution point clouds with angular resolutions around 15°, which were mostly confined to simple classification tasks [9]–[12]. More recent approaches exploit richer representations such as radar heatmaps and even raw ADC data from high-resolution radar sensors—featuring angular resolutions near 1.3° or below—to support advanced tasks like object detection, pose estimation, and segmentation [13]–[17].

In particular, the Radar Detection Transformer (RETR) [18] uses the Detection Transformer (DETR) [19] framework to fuse multi-view radar features with learnable object queries. In this framework, the tunable positional encoding (TPE) can be tuned to prioritise depth embeddings. Furthermore, this TPE has been replaced by the Differentiable Positional Encoding (DiPE) [20] scheme to automatically adjust the balance between depth and angular embeddings for bounding box (BBox) prediction. While DiPE has proven effective in refining radar-based object detection, its design has been limited to BBox estimation, thereby limiting its application to tasks that require more detailed scene understanding.

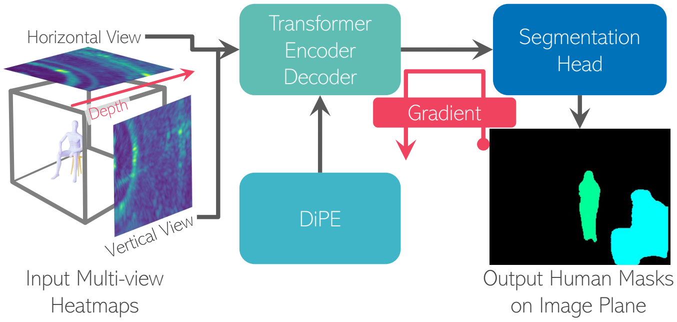
Horizontal View
Depth
Vertical View
Input Multi-view Heatmaps
Transformer Encoder Decoder
Gradient
DiPE
Segmentation Head
Output Human Masks on Image Plane

Fig. 1. End-to-End Radar human segmentation: It receives the multi-view radar heatmaps and is sent to the transformer together with DiPE, which has learnable parameters and can adjust the importance of the spatial axis. The segmentation head, which receives the decoder output, predicts the human mask. Each module is fully differentiable and is trained end-to-end.

In this paper, we propose an enhanced end-to-end segmentation pipeline that extends the DiPE framework to perform instance segmentation of human directly from radar signals. Unlike the original DiPE used in RETR for BBox regression, our approach integrates segmentation-specific mechanisms to accurately predict instance segmentation masks as illustrated in Fig. 1. By employing an extended DiPE scheme, our method dynamically modulates the relative contributions of depth and angle information during training, ensuring finer extraction of semantic features essential for segmentation tasks. Furthermore, our pipeline benefits from end-to-end optimization with a dedicated segmentation loss, thereby providing robust performance in diverse indoor environments. Evaluation on the MMVR dataset [17] indicates that our proposed segmentation pipeline achieves improved segmentation accuracy compared to conventional methods.

II. Problem Formulation

A. Multi-View Radar Heatmaps

Multi-View Radar Heatmaps are generated from raw data captured by two radar arrays: a vertical linear array and a horizontal one, as illustrated in Fig. 2. By sampling multiple reflected pulses across the array elements, a 3D raw data cube is constructed for each array, organized along ADC (intra-pulse)

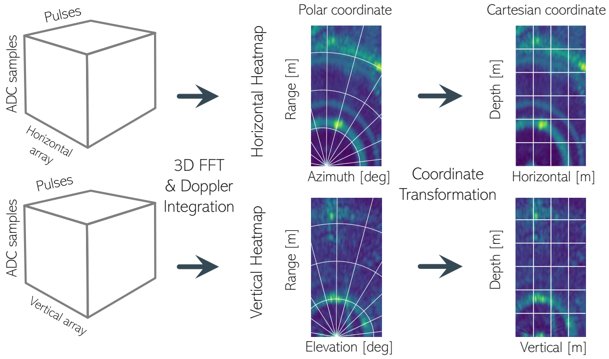

Fig. 2. Generation of multi-view heatmaps from raw radar data.

samples, pulse (inter-pulse) samples, and array elements. A 3D fast Fourier transform (FFT) converts the data cube into corresponding 3D radar spectra across the range, Doppler velocity, and spatial angle (azimuth for the horizontal array and elevation for the vertical one). To enhance the signal-to-noise ratio (SNR), the 3D radar spectra are integrated along the Doppler domain, generating two 2D radar heatmaps (range-azimuth and range-elevation) in the polar coordinate system. These heatmaps are then transformed into the radar Cartesian coordinate system, where $\mathbf{Y}_{\texttt{hor}}(m) \in \mathcal{R}^{W\times D}$ represents the horizontal-depth radar heatmap and $\mathbf{Y}_{\texttt{ver}}(m) \in \mathcal{R}^{H\times D}$ the vertical-depth heatmap for the m-th frame where W, H and D denote the number of cells of width (horizontal), height (vertical) and depth, respectively. To incorporate temporal information, M consecutive radar frames are grouped together as $\mathbf{Y}_{\texttt{hor}} \in \mathcal{R}^{M\times W\times D}$ and vertical $\mathbf{Y}_{\texttt{ver}} \in \mathcal{R}^{M\times H\times D}$.

B. Radar-Based Human Instance Segmentation

Given the multi-view radar heatmaps $\mathbf{Y}_{\texttt{hor}}$ and $\mathbf{Y}_{\texttt{ver}}$, our goal is to segment human subjects in the image plane by leveraging both heatmaps as inputs,

$$\mathbf{F}_{\texttt{image}} = \texttt{proj}_{\texttt{image}}\left(\mathcal{T}\left(f\left(\mathbf{Y}_{\texttt{hor}}, \mathbf{Y}_{\texttt{ver}}\right)\right)\right), \tag{1}$$

where $\mathbf{F}_{\texttt{image}}$ denotes segmentation masks (pixels) in the image plane, f denotes the object detection module in the radar coordinate system, $\mathcal{T}$ denotes the radar-to-camera coordinate transformation, and $\texttt{proj}_{\texttt{image}}$ denotes the 3D-to-2D image projection.

III. End-to-End Radar Human Segmentation

The end-to-end radar segmentation pipeline is illustrated in Fig. 3. It extends the radar detection transformer (RETR) [18] by incorporating optimally tuned positional encoding for human segmentation tasks. Specifically, it consists of transformer encoder and decoder modules that include self-/cross-attention, a tunable positional encoding, a radar-to-camera geometry transformation, and a 3D-to-2D projection. Since the multi-view radar features lack positional information and the self-attention is permutation-invariant, positional encoding is concatenated to the context (feature) embeddings at the input of each encoder and decoder layer. The DiPE further optimizes the tunable ratio between the angular positional embedding dimension and the depth positional embedding dimension for the human segmentation task. In the following, we introduce each module in detail.

A. Backbone

Taking the two radar heatmaps $\mathbf{Y}_{\texttt{hor}}$ and $\mathbf{Y}_{\texttt{ver}}$ as inputs, a backbone network (e.g., ResNet [21]) generates horizontal-view and vertical-view radar feature maps separately:

$$\begin{aligned} \mathbf{Z}_{\texttt{hor}} &= \texttt{backbone}\left(\mathbf{Y}_{\texttt{hor}}\right), \\ \mathbf{Z}_{\texttt{ver}} &= \texttt{backbone}\left(\mathbf{Y}_{\texttt{ver}}\right), \end{aligned} \tag{2}$$

where learnable parameters in the $\texttt{backbone}$ are shared across both views. Each feature map is generated as L multi-scale feature maps in $\mathbb{R}^{C\times \frac{W}{sl}\times\frac{D}{sl}}$ or $\mathbb{R}^{C\times \frac{H}{sl}\times\frac{D}{sl}}$ by using feature pyramid network where C, s and $l \in \{1, \cdots, L\}$ represent the number of channels, downsampling ratio over the spatial dimension and the pyramid level, respectively.

B. Top-K Selection

Since the transformer encoder expects a sequence of features as input, we map the above feature maps into a sequence of $2K$ multi-view radar features:

$$\begin{aligned} \mathbf{H}_{\texttt{hor}} &= \texttt{Selector}\left(\mathbf{Z}_{\texttt{hor}}\right) \in \mathbb{R}^{C\times K} \\ \mathbf{H}_{\texttt{ver}} &= \texttt{Selector}\left(\mathbf{Z}_{\texttt{ver}}\right) \in \mathbb{R}^{C\times K}, \end{aligned} \tag{3}$$

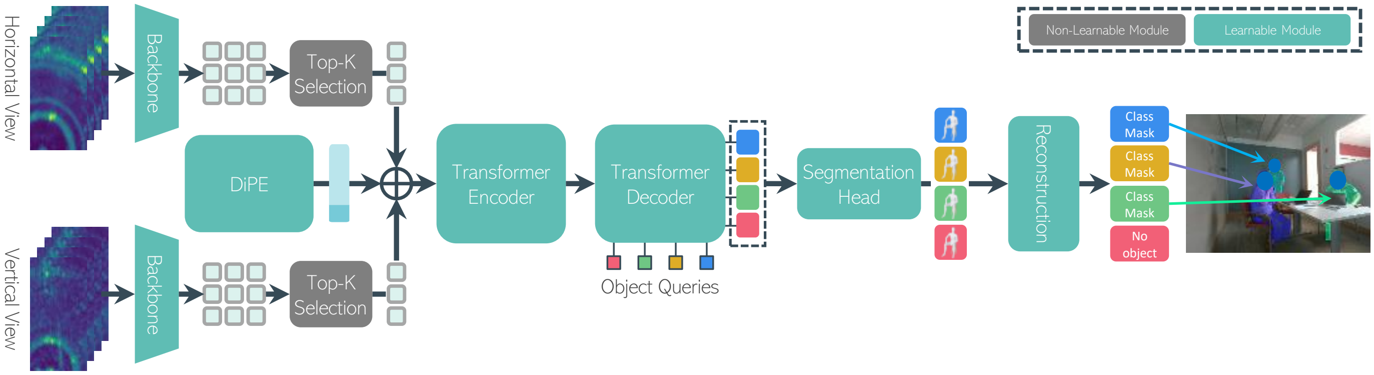

Fig. 3. The end-to-end **radar human segmentation pipeline** with differentiable positional encoding (DiPE): 1) **Encoder**: Top-K features selection and DiPE to assist feature association across the two radar views; 2) **Decoder**: the association between object queries and multi-view radar features; 3) **Segmentation Head**: enhanced object queries are enforced to estimate binary pixels as the segmentation masks of human subjects in the image plane.

where $K \ll \min\{WD/s^2, HD/s^2\}$. We further pool these $2K$ multi-view radar features

$$\mathbf{H}^0 = [\mathbf{H}_{\mathtt{hor}}, \mathbf{H}_{\mathtt{ver}}] \in \mathbb{R}^{C \times 2K} \tag{4}$$

as the input to the transformer encoder module.

C. Transformer Encoder and Decoder

The $2K$ multi-view radar features can be considered as a sequence of $2K$ input tokens for the transformer encoder. For the l-th encoder layer, we have

$$\mathbf{H}^{l+1} = \mathtt{encoder}\left(\mathbf{H}^l\right), \quad l = 0, 1, \cdots, L_{\mathtt{self}} - 1. \tag{5}$$

The encoder is mainly to process the multi-view radar input sequences (i.e., $2K$ selected radar features) and generate rich contextual representations by leveraging self-attention and feed-forward transformations and capture feature correspondence across the two radar views.

For each decoder layer, it takes N object queries: $\mathbf{Q}^l \in \mathbb{R}^{C \times N}$ as its input, and consists of a self-attention layer, a cross-attention layer and a feed forward network (FFN). It updates all queries

$$\bar{\mathbf{Q}}^l = \mathtt{decoder}_{\mathtt{self}}\left(\mathbf{Q}^l\right), \tag{6}$$

$$\mathbf{Q}^{l+1} = \mathtt{decoder}_{\mathtt{cross}}\left(\bar{\mathbf{Q}}^l, \mathbf{H}^{L_{\mathtt{self}}}\right), \tag{7}$$

via multi-head self-attention and cross-attention, respectively, where $l = 0, 1, \cdots, L_{\mathtt{cross}} - 1$.

D. Differentiable Positional Encoding (DiPE)

Positional encoding is essential for providing spatial information to each feature (embedding $\mathbf{h} \in \mathbf{H}^l$ or decoder embedding $\mathbf{q} \in \mathbf{Q}^l$). We adopt DiPE as the positional encoding, leveraging the shared depth axis between the two radar views as an inductive bias to emphasize depth importance and reduce redundant correlations. DiPE first generates the positional embeddings of dimension $d_{\mathtt{pos}}$ for each axis (vertical, horizontal, depth) in advance. Then, using the parameters $\boldsymbol{\theta}$, we generate a mask $m\,(i; \boldsymbol{\theta})$ and apply the dual masking:

$$\mathbf{p} = \mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta}) \odot \mathbf{d} + (\mathbf{1} - \mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta})) \odot \mathbf{a}_{\mathtt{f}}, \tag{8}$$

where $\mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta}) = \left\{m\,(1; \boldsymbol{\theta}), ..., m\left(d_{\mathtt{pos}}; \boldsymbol{\theta}\right)\right\}^{\top}$ is the vector collected with each dimension i, $\mathbf{1}$ is a vector with all elements of 1, $\odot$ represents Hadamard product, and $\mathtt{f}$ is an operation that flips the order of the vector's elements: $\mathbf{a}_{\mathtt{f}}^{(i)} = \mathbf{a}^{(d_{\mathtt{pos}}+1-i)}$. And the mask $m\,(i; \boldsymbol{\theta})$ is specifically defined to be differentiable [20]:

$$m\,(i; \boldsymbol{\theta}) = 1 - \frac{1}{1 + e^{-\tau(i-\mu)}}, \tag{9}$$

where $\boldsymbol{\theta} = \{\mu \geq 0, \tau\}$ are offset and temperature parameters.

The attention weight is based on the dot-product between query ($\mathtt{q}$) and key ($\mathtt{k}$):

$$\begin{aligned}
&\left(\mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta}) \odot \mathbf{d}_{\mathtt{q}} + (\mathbf{1} - \mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta})) \odot \mathbf{a}_{\mathtt{f},\mathtt{q}}\right)^{\top} \\
&\qquad \left(\mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta}) \odot \mathbf{d}_{\mathtt{k}} + (\mathbf{1} - \mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta})) \odot \mathbf{a}_{\mathtt{f},\mathtt{k}}\right) \\
&= \left(\bar{\mathbf{d}}_{\mathtt{q}} + \mathbf{a}_{\mathtt{f},\mathtt{q}} - \bar{\mathbf{a}}_{\mathtt{f},\mathtt{q}}\right)^{\top} \left(\bar{\mathbf{d}}_{\mathtt{k}} + \mathbf{a}_{\mathtt{f},\mathtt{k}} - \bar{\mathbf{a}}_{\mathtt{f},\mathtt{k}}\right) \\
&= \bar{\mathbf{d}}_{\mathtt{q}}^{\top} \bar{\mathbf{d}}_{\mathtt{k}} + \bar{\mathbf{a}}_{\mathtt{f},\mathtt{q}}^{\top} \bar{\mathbf{a}}_{\mathtt{f},\mathtt{k}} + \mathbf{a}_{\mathtt{f},\mathtt{q}}^{\top} \bar{\mathbf{d}}_{\mathtt{k}} - \bar{\mathbf{a}}_{\mathtt{f},\mathtt{q}}^{\top} \bar{\mathbf{d}}_{\mathtt{k}} + \bar{\mathbf{d}}_{\mathtt{q}}^{\top} \mathbf{a}_{\mathtt{f},\mathtt{k}} \\
&\qquad - \bar{\mathbf{d}}_{\mathtt{q}}^{\top} \bar{\mathbf{a}}_{\mathtt{f},\mathtt{k}} + \mathbf{a}_{\mathtt{f},\mathtt{q}}^{\top} \mathbf{a}_{\mathtt{f},\mathtt{k}} - \mathbf{a}_{\mathtt{f},\mathtt{q}}^{\top} \bar{\mathbf{a}}_{\mathtt{f},\mathtt{k}} - \bar{\mathbf{a}}_{\mathtt{f},\mathtt{q}}^{\top} \mathbf{a}_{\mathtt{f},\mathtt{k}},
\end{aligned} \tag{10}$$

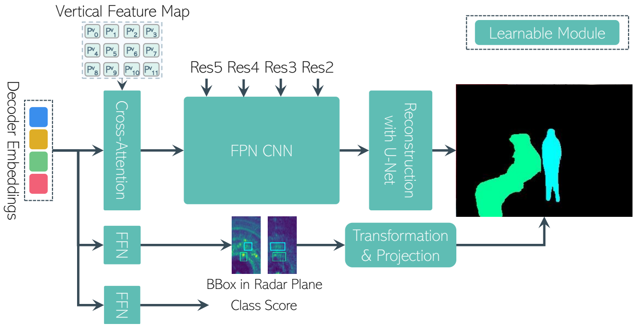

Fig. 4. Illustration of the segmentation head and reconstruction modules.

where $\bar{\mathbf{x}} = \mathbf{m}_{\mathtt{dual}}\,(\boldsymbol{\theta}) \odot \mathbf{x}$. Eq. 10 contains blended components according to τ.

E. Segmentation Head and Reconstruction

We extend RETR for instance segmentation by incorporating a dedicated segmentation head that operates on the decoder outputs. Fig. 4 illustrates the overall design, consisting of four main components: 1) a cross-attention layer, 2) an FPN-style CNN, and 3) a lightweight U-Net [22], and 4) BBox and Class prediction modules. Each refined query $\mathbf{q} \in \mathbf{Q}^{L_{\mathtt{cross}}}$ is processed through a cross-attention layer to generate low-resolution attention heatmaps for individual objects.

$$\mathbf{z} = \mathtt{Att}_{\mathtt{cross}}\,(\mathbf{q}, \mathbf{Z}_{\mathtt{ver}}), \tag{11}$$

where $\mathbf{Z}_{\mathtt{ver}}$ is the vertical feature map, which emphasizes the subject's height dimension, and is fed into a feature pyramid network (FPN) $\mathtt{FPN}(\cdot)$. By aggregating multi-scale vertical features from layers (Res5 to Res2), the FPN not only refines the resolution but also handles the transformation from the radar plane to the image plane. However, since the FPN alone has limited capacity to produce high-resolution masks, we append a light U-Net $\mathtt{Unet}(\cdot)$ as the reconstruction to further enhance the coarse masks:

$$\hat{\mathbf{m}} = \mathtt{sigmoid}\,(\mathtt{Unet}\,(\mathtt{FPN}\,(\mathbf{z}, \mathbf{Z}_{\mathtt{hor}}, [\mathbf{Z}_{\mathtt{ver}}]))) \tag{12}$$

where $\hat{\mathbf{m}}$ is a predicted binary mask and $[\mathbf{Z}_{\mathtt{ver}}]$ denote the multi-scale vertical feature maps obtained by Backbone. Our model outputs a single binary mask for each query with above sigmoid function. Following the BBox prediction in the radar domain, we apply the Radar-to-Camera transformation and project the 3D box onto the image plane. The corresponding image-region ground-truth (GT) mask is then cropped and used to supervise the predicted mask for that specific query.

F. Loss Function

We adopt a combination of Dice/F-1 loss [23] and focal loss [24] for training the segmentation head. Let $m_p \in [0, 1]$ be the GT label and $\hat{m}_p \in [0, 1]$ in $\{\hat{\mathbf{m}}_i\}_{i=1}^{N}$ be the predicted probability for pixel p. Dice loss and focal loss are defined as:

$$\mathcal{L}_{\mathtt{dice}} = 1 - \frac{2\sum_p m_p\,\hat{m}_p}{\sum_p m_p + \sum_p \hat{m}_p + \epsilon}, \tag{13}$$

$$\begin{aligned}
\mathcal{L}_{\mathtt{focal}} = &-\sum_p \{\alpha\, m_p (1 - \hat{m}_p)^{\gamma} \log(\hat{m}_p) \\
&+ (1 - \alpha)\,(1 - m_p)\,(\hat{m}_p)^{\gamma} \log(1 - \hat{m}_p)\},
\end{aligned} \tag{14}$$

TABLE I

Detection results on MMVR. RFMask does not have PE. For RETR with TPE, `Trained` $\boldsymbol{\theta}$ corresponds to a ratio α.

`Model`	`PE`	`Trained` $\boldsymbol{\theta} = \{\mu, \tau\}$	`AP`	`AP`$_{50}$	`AP`$_{75}$	`AR`$_{1}$	`AR`$_{10}$	`IoU`
RFMask	-	-	31.37	61.50	27.48	33.23	38.41	65.30
DETR	`Sinusoid`	-	29.38	62.31	25.35	31.32	43.06	70.15
RETR with TPE	`Sinusoid`	0.60 / -	46.75	83.80	46.06	42.19	57.39	77.21
RETR with TPE	`Learned`	0.60 / -	46.71	82.27	45.09	41.61	56.22	76.1
RETR with DiPE (Ours)	`Sinusoid`	0.90 / 0.94	47.09	84.15	46.14	44.43	59.18	77.55
RETR with DiPE (Ours)	`Learned`	0.67 / 0.32	47.75	83.72	46.31	42.11	56.37	77.01

where ϵ is a small constant for numerical stability, α is a balancing parameter that adjusts the importance between positive and negative samples, and γ is a focusing parameter that down-weights the loss contribution of well-classified examples. The final segmentation loss is the sum of these two terms:

$$\mathcal{L}_{\texttt{seg}} = \mathcal{L}_{\texttt{dice}} + \mathcal{L}_{\texttt{focal}}. \tag{15}$$

IV. Experiments

A. Experimental Setup

Our experiments utilize the indoor radar dataset MMVR [17], following the same protocol as in [18], [20]. MMVR comprises multi-view radar heatmaps collected from 25 human subjects across 6 rooms over 9 days. In this work, we focus on data from Protocol 2 (P2), which contains a total of 237.9K frames capturing multiple subjects. The training, validation, and test splits are defined according to the S1 split provided in the MMVR dataset.

Two different positional encoding strategies are investigated: a sine/cosine encoding method, denoted as `Sinusoid`, and a learnable embedding approach, denoted as `Learned`. For baseline comparisons, we consider RFMask [16], DETR [19], and RETR [18]. Specifically, in the implementations of DETR and RETR, the embedding dimension is set to $d_{\texttt{pos}} = 256$, and a ratio of $\alpha = 0.6$ is employed as established in [18]. All other hyper-parameters, including the learning rate and early stopping criteria, are adopted from [18].

Performance is evaluated from multiple perspectives. For segmentation, we assess instance segmentation performance by evaluating the Intersection over Union (`IoU`) [25] between predicted masks and GT masks. `IoU` is the ratio of the overlap to the union of a predicted BBox A and annotated BBox B as:

$$\texttt{IoU}(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|}. \tag{16}$$

B. Results

Table I presents a comprehensive comparison of detection models, where the detection results (AP metrics) are directly quoted from Table 1 in [20]. In the table, "`PE`" indicates the type of positional encoding used, and "`Trained` $\boldsymbol{\theta} = \{\mu, \tau\}$" represents the learned parameters obtained after training. For RETR with TPE, the value of the ratio α is reported as μ, whereas for DiPE, both the pre-scaling value μ and the blending parameter τ are provided. The segmentation performance, evaluated using the `IoU` metric, further supports the benefits of DiPE. The higher `IoU` scores achieved by DiPE compared to the baseline models indicate that our approach not only improves detection accuracy but also yields more precise instance segmentation by ensuring a better overlap between the predicted masks and GT.

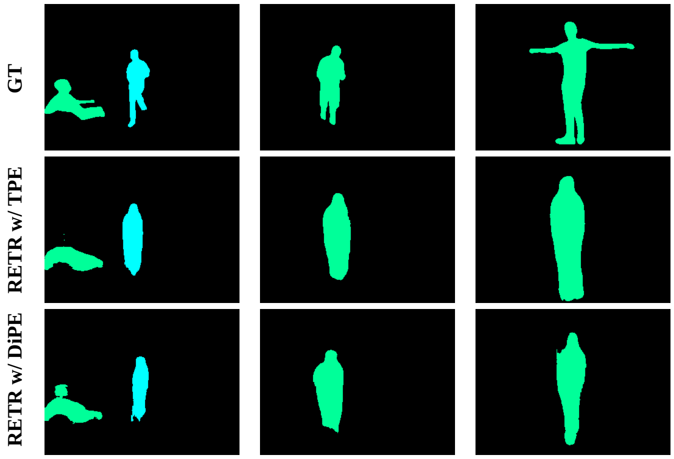

Fig. 5. Visualization of human segmentation in the image plane. Each row represents a different environment. GT denotes ground-truth for comparison of predicted masks. Our RETR w/ DiPE can predict the human shape more accurately.

Fig. 5 shows the visualization results of the segmentation. For RETR w/ TPE, the segmentation results are significantly distorted, with human shapes appearing overly smoothed or stretched, losing important shape details. On the other hand, our RETR w/ DiPE is more accurate and preserves the human shape better compared to RETR w/ TPE. This shows that our RETR w/ DiPE method provides a more accurate and realistic human segmentation, outperforming RETR w/ TPE in preserving the correct shape and details of human figures.

V. Conclusion

We introduced an end-to-end radar human segmentation pipeline that takes the multi-view radar heatmaps as input and estimates binary masks for each human subject in the image plane. Our pipeline leverages the query-based detection frameworks such as DETR and RETR and integrates the differentiable positional encoding to enhance the segmentation performance. Evaluations on the MMVR dataset confirm the effectiveness of the proposed pipeline.

References

[1] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and

challenges," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 98–117, 2020.

[2] A. Pandharipande, C.-H. Cheng, J. Dauwels, S. Z. Gurbuz, J. Ibanez-Guzman, G. Li, A. Piazzoni, P. Wang, and A. Santra, "Sensing and machine learning for automotive perception: A review," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 11 097–11 115, 2023.

[3] R. Yataka, P. Wang, P. Boufounos, and R. Takahashi, "Radar perception with scalable connective temporal relations for autonomous driving," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13 266–13 270.

[4] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 050–17 059.

[5] R. Yataka, P. Wang, P. Boufounos, and R. Takahashi, "SIRA: Scalable inter-frame relation and association for radar perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15 024–15 034.

[6] M. Amin, *Radar for Indoor Monitoring: Detection, Classification, and Assessment*. CRC Press, 2017.

[7] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "RF-based 3D skeletons," in *Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 267–281.

[8] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[9] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 4100–4109.

[10] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[11] P. Gong, C. Wang, and L. Zhang, "MMPoint-GNN: Graph neural network with dynamic edges for human activity recognition through a millimeter-wave radar," in *International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–7.

[12] F. Jin, A. Sengupta, and S. Cao, "mmFall: Fall detection using 4-D mmWave radar and a hybrid variational RNN autoencoder," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1245–1257, 2022.

[13] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941–3952, 2018.

[14] G. Li, Z. Zhang, H. Yang, J. Pan, D. Chen, and J. Zhang, "Capturing human pose using mmWave radar," in *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020, pp. 1–6.

[15] S. Yang and Y. Kim, "Single 24-GHz FMCW radar-based indoor device-free human localization and posture sensing with cnn," *IEEE Sensors Journal*, vol. 23, no. 3, pp. 3059–3068, 2023.

[16] Z. Wu, D. Zhang, C. Xie, C. Yu, J. Chen, Y. Hu, and Y. Chen, "RFMask: A simple baseline for human silhouette segmentation with radio signals," *IEEE Transactions on Multimedia*, vol. 25, pp. 4730–4741, 2023.

[17] M. M. Rahman, R. Yataka, S. Kato, P. Wang, P. Li, A. Cardace, and P. Boufounos, "MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception," in *Computer Vision – ECCV 2024*. Cham: Springer Nature Switzerland, 2025, pp. 306–322. [Online]. Available: https://zenodo.org/records/12611978

[18] R. Yataka, A. Cardace, P. P. Wang, P. Boufounos, and R. Takahashi, "RETR: Multi-view radar detection transformer for indoor perception," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://github.com/merlresearch/radar-detection-transformer

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–-229.

[20] R. Yataka, P. P. Wang, P. Boufounos, and R. Takahashi, "Multi-view radar detection transformer with differentiable positional encoding," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

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

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[23] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. [Online]. Available: https://doi.org/10.1109/ICCV.2017.324

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.