milliMamba: Specular-Aware Human Pose Estimation via Dual mmWave Radar with Multi-Frame Mamba Fusion

Niraj Prakash Kini[†], Shiau-Rung Tsai[†], Guan-Hsun Lin[†],
Wen-Hsiao Peng[†], Ching-Wen Ma[†], Jenq-Neng Hwang[‡]
[†]National Yang Ming Chiao Tung University, Taiwan, [‡]University of Washington, USA
{nirajnycu.ee06, mick20001108.cs12, abc900203abc.cs12}@nycu.edu.tw,
wpeng@cs.nycu.edu.tw, machingwen@nycu.edu.tw, hwang@uw.edu

Poster Booth: 07 (March-08, 16:00~17:45)

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

國立陽明交通大學
NYCU

Outline

- Introduction
- Main Architecture
- Experimental Results
- Conclusion

Outline

- Introduction
- Main Architecture
- Experimental Results
- Conclusion

Introduction

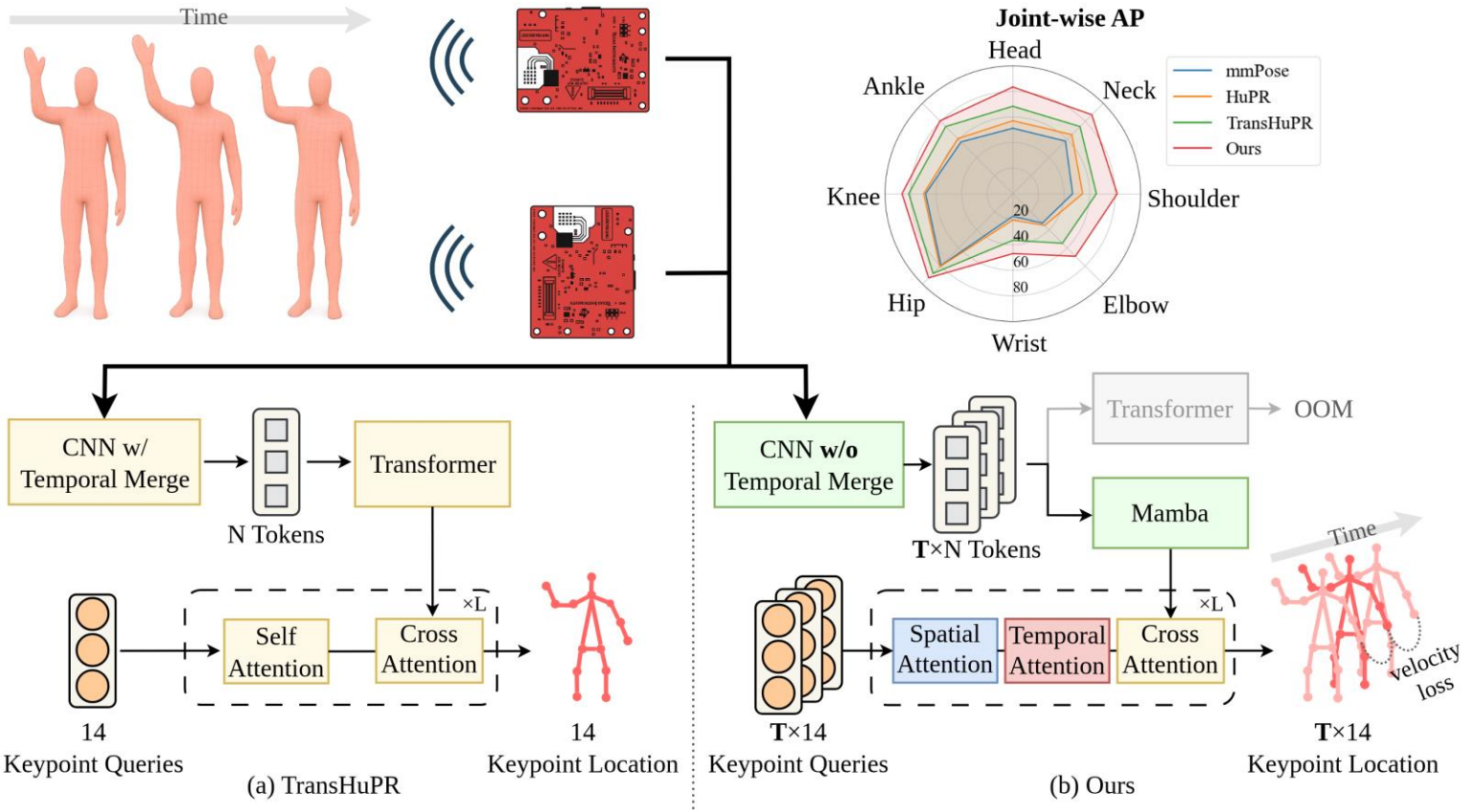
Why Radar-based Human Pose Estimation?

- RGB cameras → Privacy concerns, Sensitive to lighting
- Radar → Privacy-preserving, works in darkness, smoke, occlusion

Core Challenge: Specular Reflection

- Radar only captures surfaces reflecting back
- Small joints (wrist, ankle) are poorly reflected
- Temporal instability
- Orientation sensitivity

Previous Methods and Their Shortcomings



- Early temporal fusion → Collapse time dimension
- Many-to-one prediction → Lose temporal supervision
- Transformer quadratic complexity → Cannot scale to long sequences

Solution → Mamba

Why Mamba?

- Linear complexity → longer sequences
- Recovers missing joints via temporal context
- Models space, time, and multi-view jointly
- Higher accuracy with lower memory

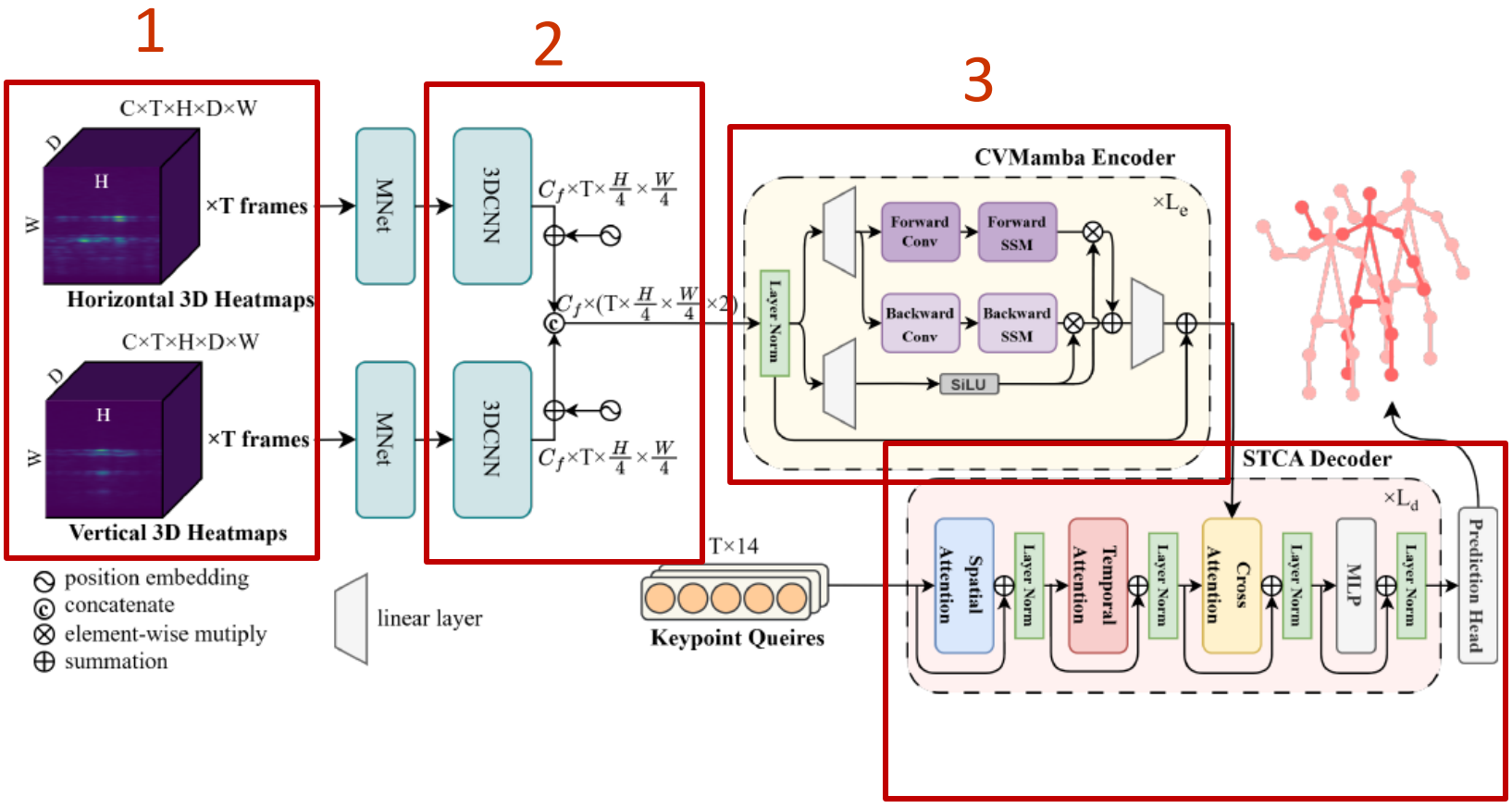
Contributions of milliMamba

- CVMamba Encoder enables long-range spatio-temporal modeling with linear complexity and cross-view fusion.
- STCA Decoder leverages multi-frame attention to improve accuracy and handle missing joints.
- milliMamba achieves state-of-the-art results on HuPR and TransHuPR benchmarks.

Main Architecture

- Introduction
- Main Architecture
- Experimental Results
- Conclusion

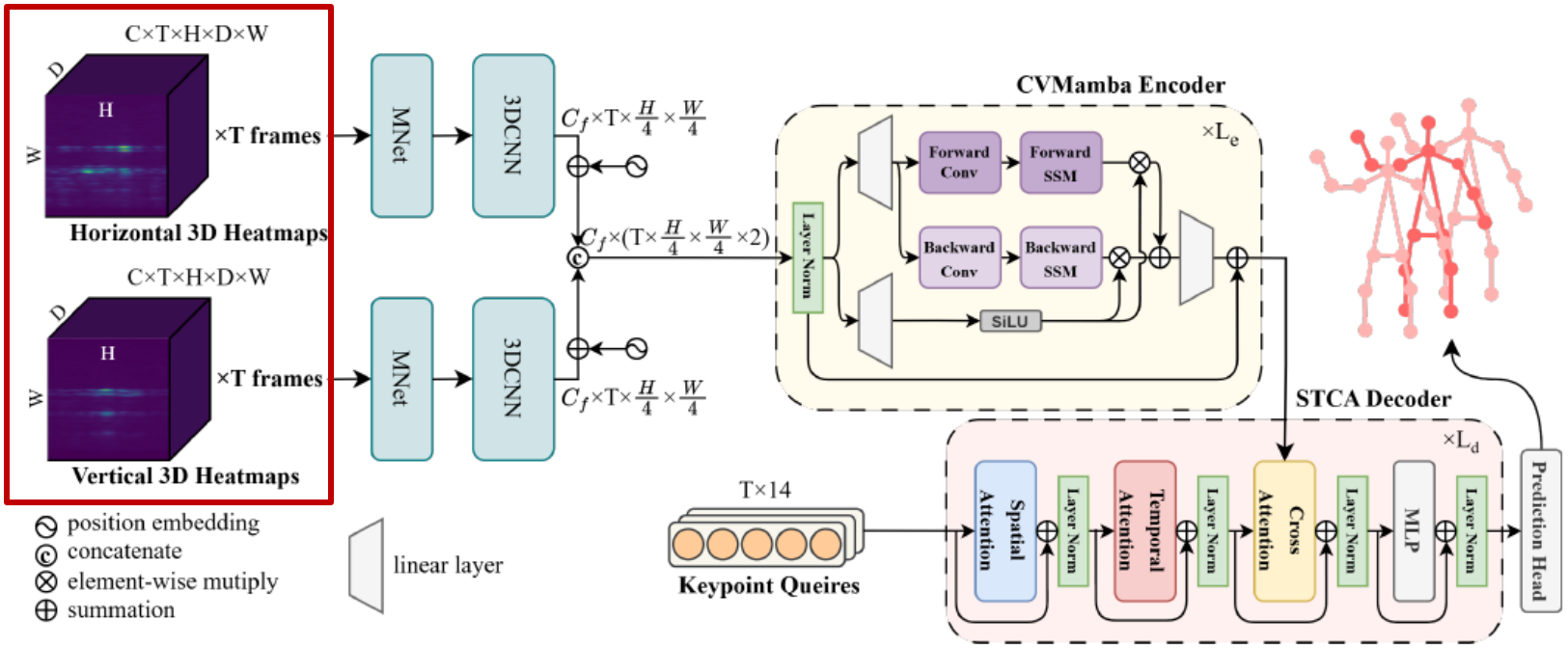
Main Architecture



1. Input Represenation
2. Cross-View Feature Fusion
3. CVMamba Encoder
4. Multi-Pose STCA Decoder

Input Representation

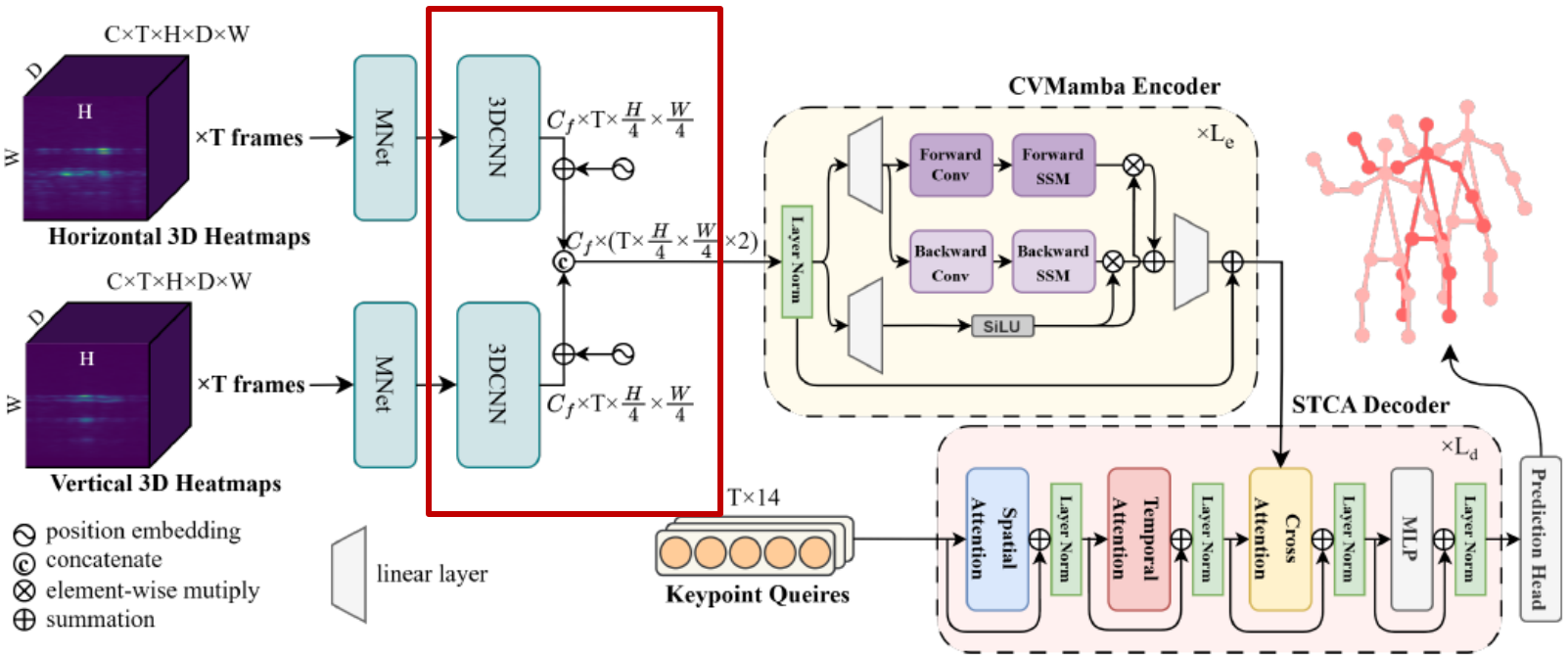
1



- Dual radar views (Horizontal + Vertical)
- 3D FFT → Angle × Doppler × Range heatmaps
- Real & Imaginary parts → 2 channels

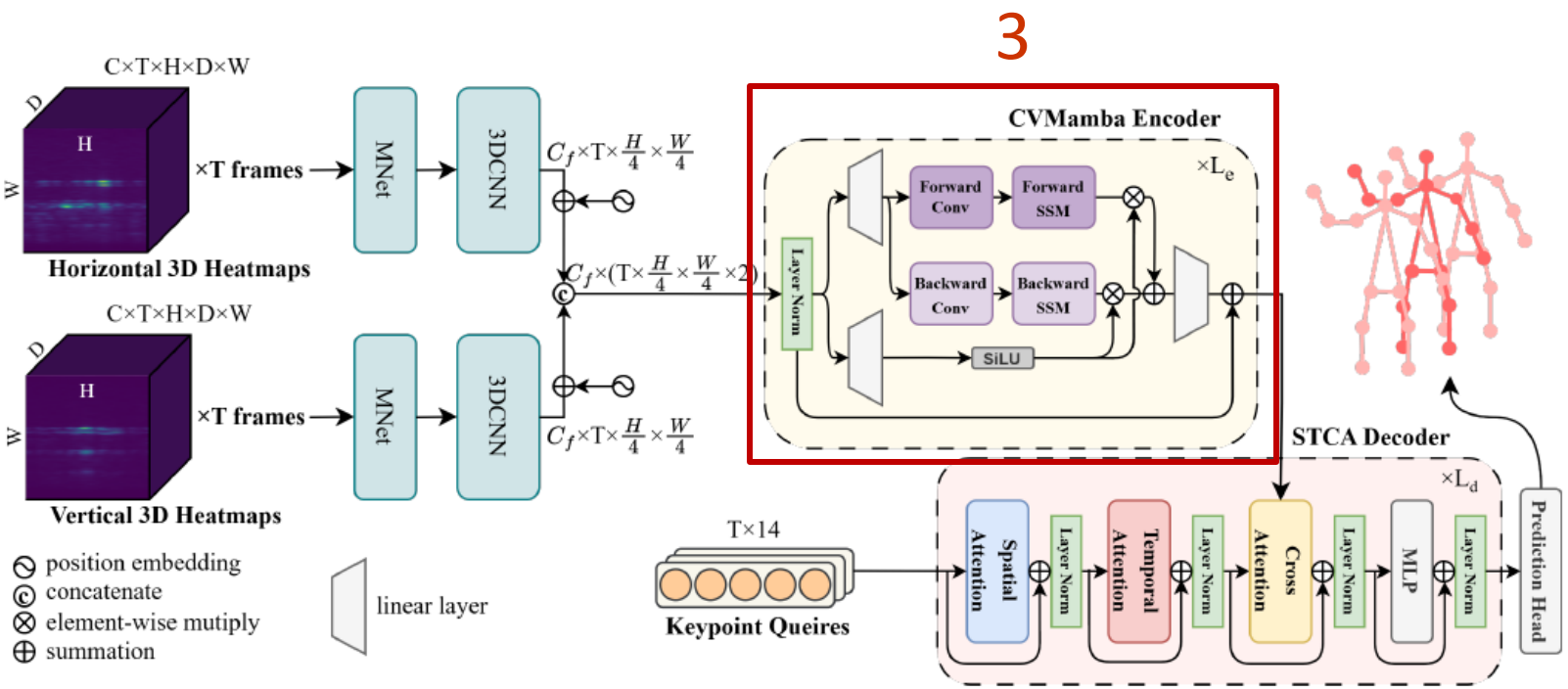
Cross-View Fusion

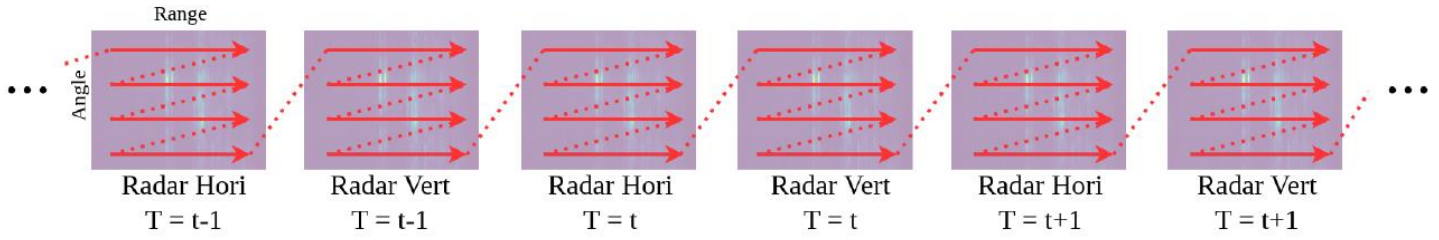
2



- Two CNN branches (per view)
- Learnable positional embeddings
- Feature concatenation across views

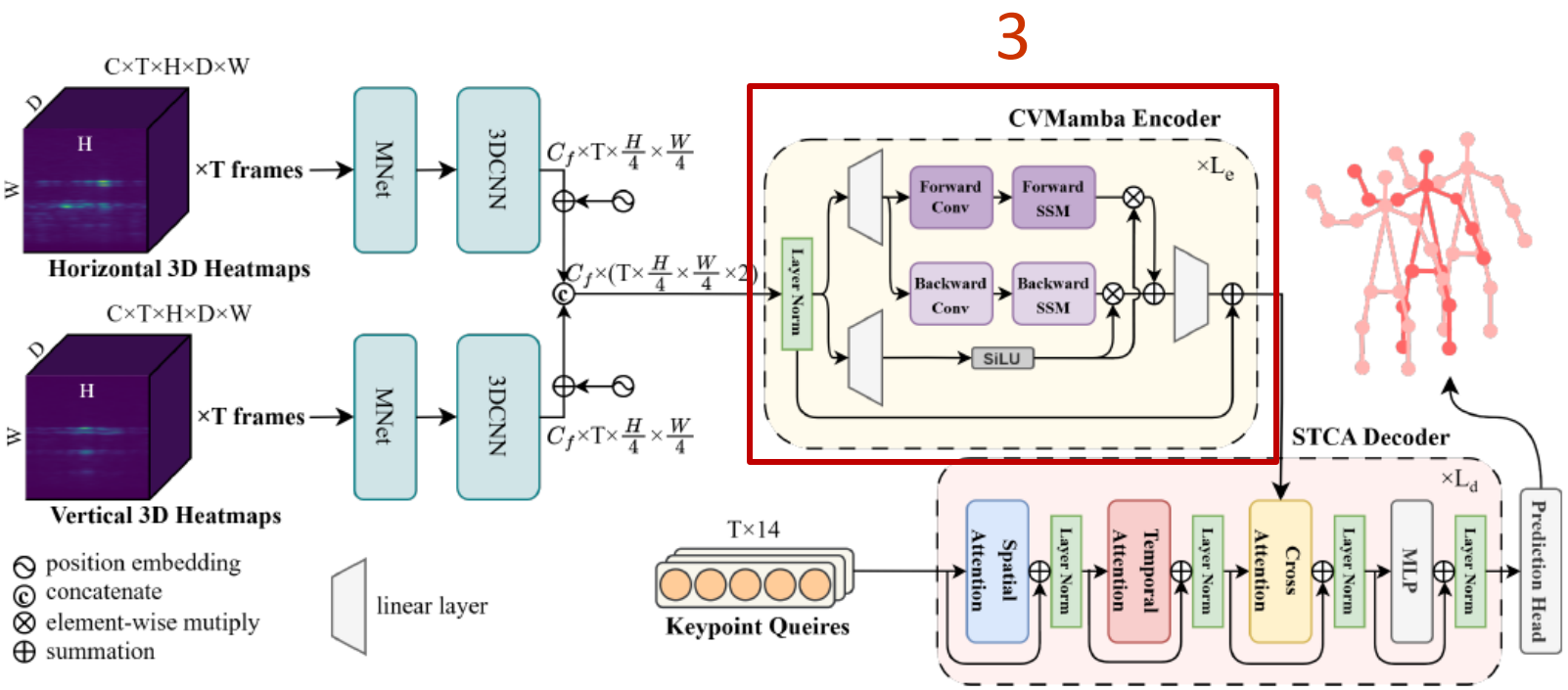
Zigzag Scanning Strategy





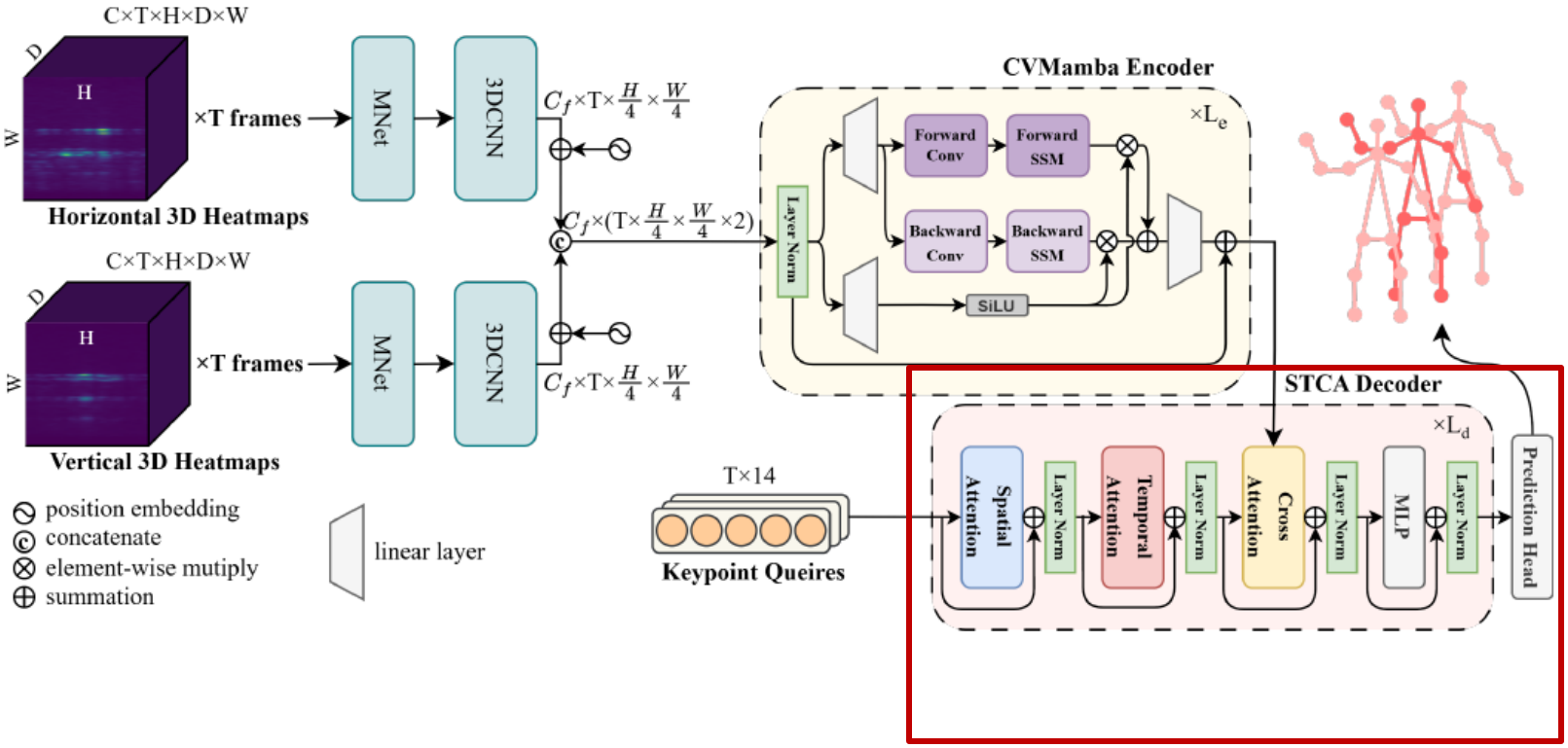
- Range → Angle → View → Frame
- Bidirectional processing

CVMamba Encoder



- State Space Model (SSM) → Linear Complexity
- Long-range spatio-temporal modeling
- Linear memory growth (vs quadratic Transformer)

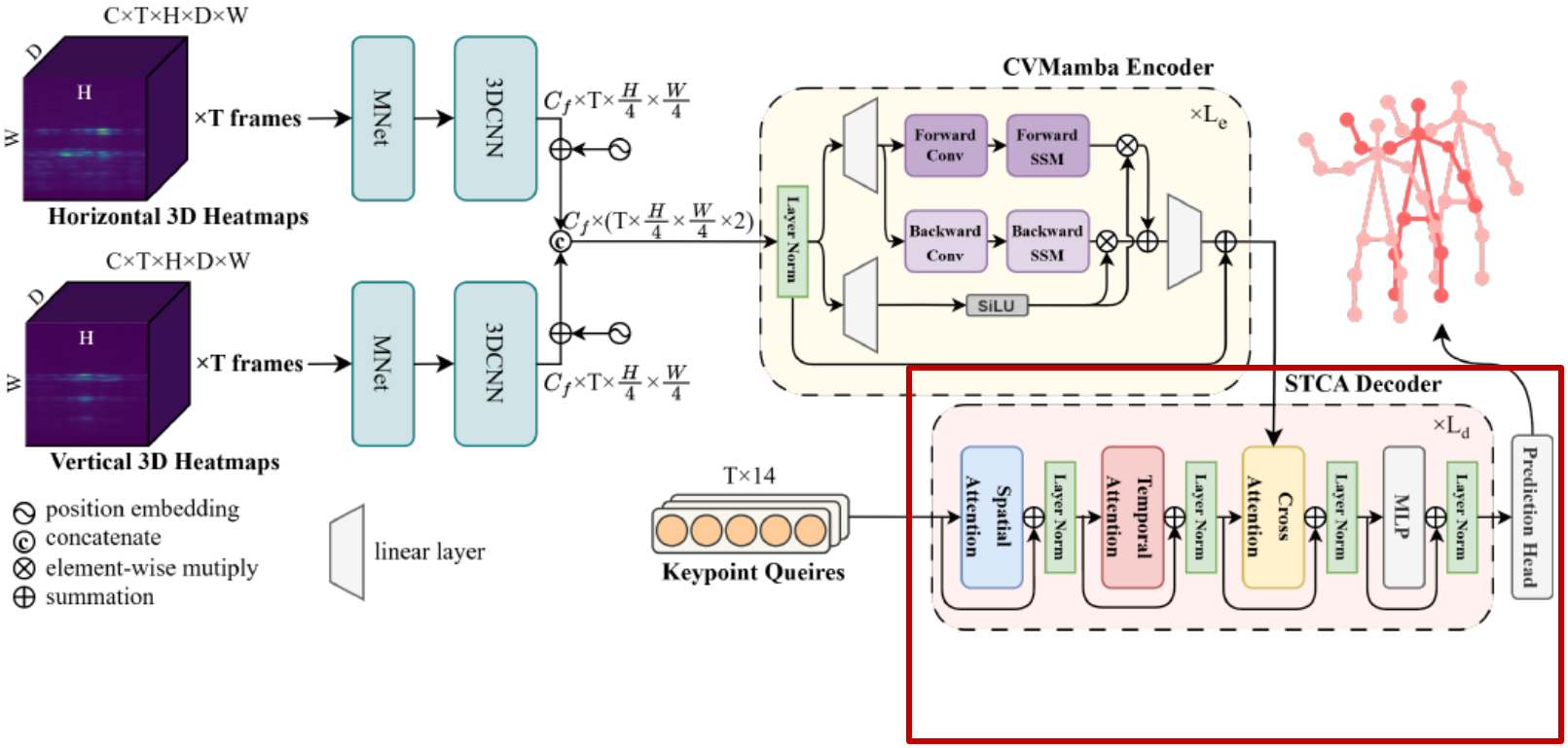
Multi-Pose STCA Decoder



4

- **Learnable Keypoint Queries**
- J × T queries (Joint × Frame)
- DETR-style decoding
- Each query = one joint at one frame

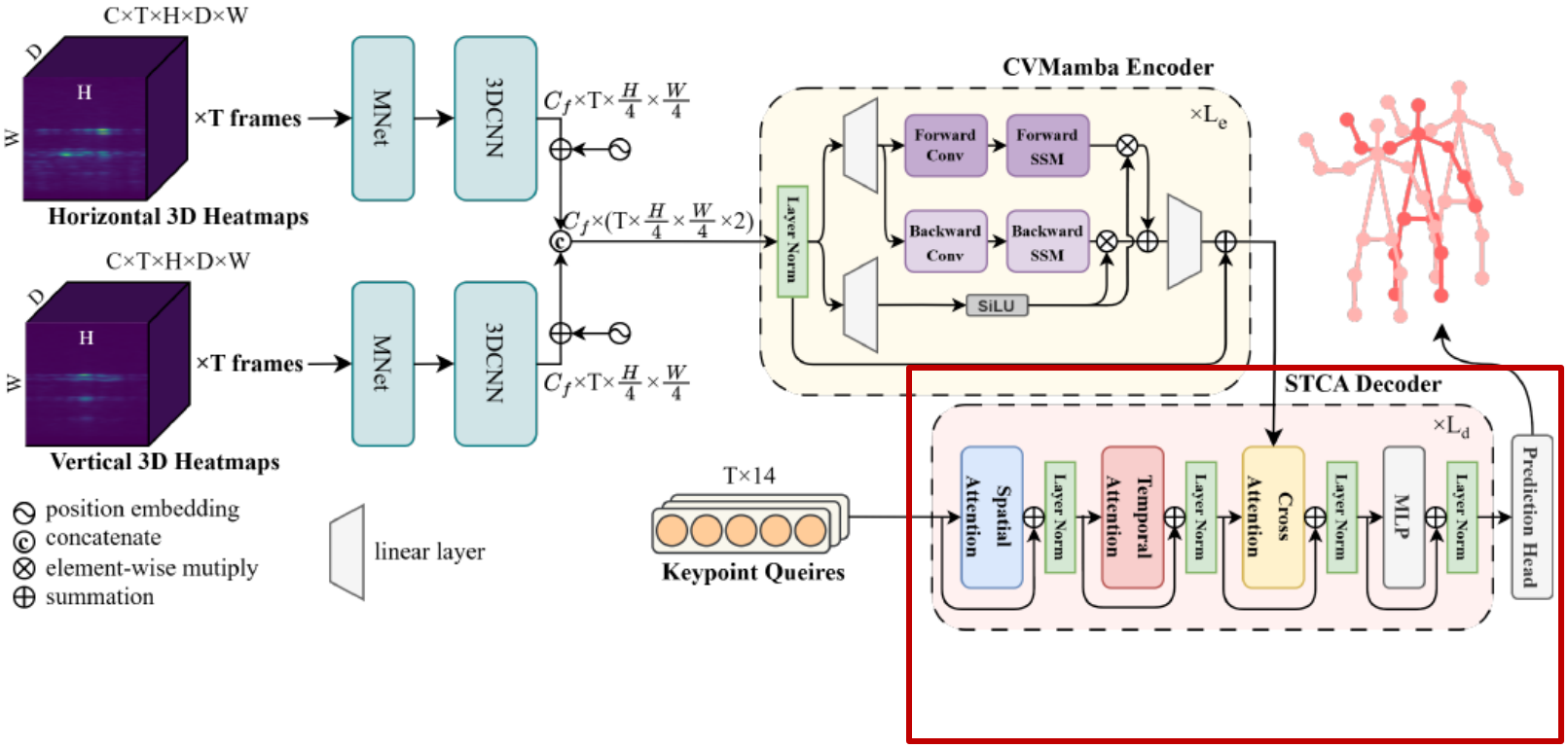
Multi-Pose STCA Decoder



4

- **Spatio-Temporal Self-Attention**
- Spatial Attention → Joint relations within frame
- Temporal Attention → Same joint across frames
  → Motion consistency

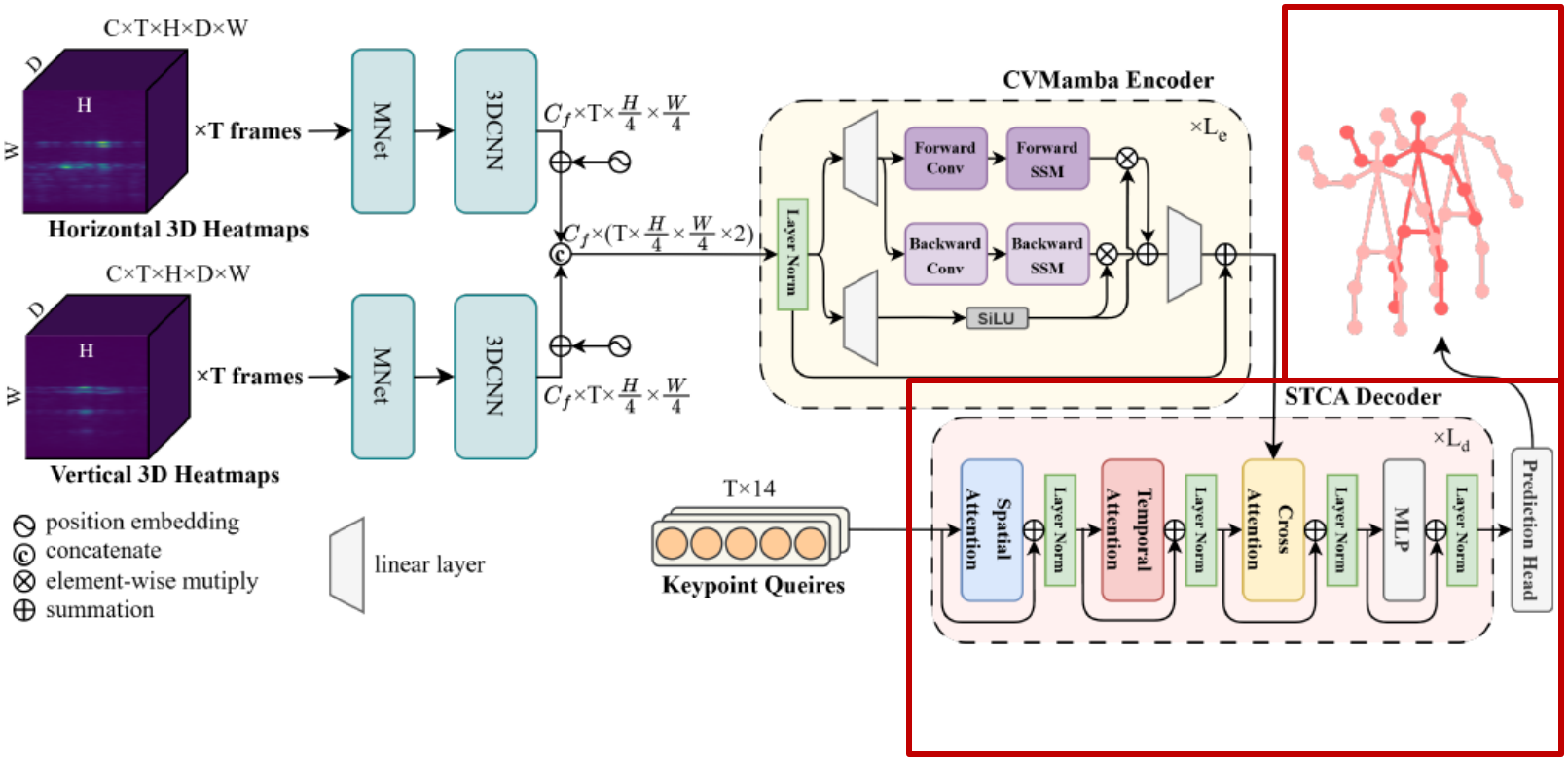
Multi-Pose STCA Decoder



4

- **Cross-Attention to Encoder**
- Queries attend to CVMamba features
- Global contextual reasoning
- Recover missing joints (specular reflection)

Multi-Pose STCA Decoder



4

- **Many-to-Many Prediction**
- Predict T poses simultaneously
- +4.1 AP over many-to-one
- Only center frame used at inference

Training Objective

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{oks}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}$$

Velocity Loss (Temporal Smoothness)

$$\mathcal{L}_{vel} = \frac{1}{(T-1)\mathcal{J}} \sum_{f=1}^{T-1} \sum_{j=1}^{J} \left\| \hat{v}_{f,j} - v_{f,j} \right\|_2^2$$

Predicted velocity of joint *j* at frame *f* as the $\hat{v}_{f,j}$ which is the difference between predicted joint and groound truth joint

Outline

- Introduction
- Main Architecture
- Experimental Results
- Conclusion

Experimental Results

- MilliMamba on TransHuPR Dataset

Method	Complexity			Joint-wise AP								Overall AP		
	MACs	Params	Mem	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	AP	AP^{50}	AP^{75}
mmPose [23]	**85.5 M**	15.0 M	**67.2 MB**	51.2	58.2	46.8	32.6	17.7	79.5	68.3	57.3	48.4	88.4	47.4
HuPR [13]	68.6 G	35.5 M	339.7 MB	57.1	65.3	54.6	35.2	20.6	80.8	69.8	60.9	51.5	89.5	53.7
TransHuPR [12]	5.8 G	5.3 M	230.8 MB	68.4	74.3	65.4	54.9	36.5	88.3	81.5	74.3	67.5	96.9	76.7
Ours	34.4 G	**4.0 M**	224.1 MB	**83.5**	**87.4**	**81.7**	**69.3**	**46.9**	**93.2**	**86.7**	**80.6**	**78.5**	**98.7**	**89.3**

- MilliMamba on HuPR Dataset

Method	Complexity			Joint-wise AP								Overall AP		
	MACs	Params	Mem	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	AP	AP^{50}	AP^{75}
mmPose [23]	**85.5 M**	15.0 M	**67.2 MB**	56.1	60.9	40.6	24.9	14.2	63.2	58.6	56.1	41.4	79.4	38.3
HuPR [13]	68.6 G	35.5 M	339.7 MB	77.5	81.9	70.3	45.5	22.3	88.1	82.2	73.1	63.4	97.0	74.0
TransHuPR [12]	5.8 G	5.3 M	230.8 MB	77.1	78.6	63.2	55.6	44.9	84.5	83.6	80.0	69.4	95.1	79.9
Ours	34.4 G	**4.0 M**	224.1 MB	**90.0**	**91.8**	**83.2**	**75.2**	**59.5**	**94.3**	**93.6**	**89.3**	**84.0**	**98.5**	**94.9**

Experimental Results

- Input Representation

Input Representation	AP	AP^{50}	AP^{75}
density map	58.5	92.5	62.7
4D FFT	72.0	97.3	81.8
3D FFT	**74.5**	**98.5**	**84.7**

- Single Pose vs Multi-Pose

Prediction Strategy	AP	AP^{50}	AP^{75}
Many-to-one	70.4	97.0	81.0
Many-to-many	**74.5**	**98.5**	**84.7**

Experimental Results

- One Radar vs Two Radars

Radar Used	Complexity			Overall AP		
	MACs	Params	Mem	AP	AP^{50}	AP^{75}
Hori	17.3 G	3.2 M	121.2 MB	67.3	95.8	75.0
Vert	17.3 G	3.2 M	121.2 MB	74.5	98.5	84.7
Hori+Vert	34.4 G	4.0 M	224.1 MB	78.5	98.7	89.3

- Transformer vs CVMamba

Encoder	Complexity			Overall AP		
	MACs	Params	Mem	AP	AP^{50}	AP^{75}
Transformer	14.9 G	3.9 M	610.3 MB	65.4	95.5	73.5
Mamba	5.6 G	3.2 M	44.7 MB	66.9	95.7	75.0

Visualization Results

Outline

- Introduction
- Main Architecture
- Experimental Results
- Conclusion

Conclusion

- milliMamba addresses sparse specular reflections in dual mmWave radar for robust 2D human pose estimation.
- It combines efficient 3D FFT preprocessing with multi-frame Mamba-based spatio-temporal modeling.
- The framework achieves state-of-the-art performance on both TransHuPR and HuPR benchmarks.
- It delivers a strong balance between accuracy and computational efficiency.

Q&A

Poster Booth: 07 (March-08, 16:00~17:45)

Visit Project Page