TARSS-Net: Temporal-Aware Radar Semantic Segmentation Network

(#9831)

YOUCHENG ZHANG*, LIWEN ZHANG*,
TENG LI, ET AL.
DEC. 2024.

Contents.

I. BACKGROUND	II. MOTIVATION	III. METHODOLOGY	VI. RESULTS
☐ **Deep discussion and analysis of current temporal modeling paradigm**	☐ **Design principles of spatio-temporal encoding for Radar Semantic Segmentation**	☐ **Temporal Relation Attentive Model (TRAM)** ◆ Target-History Temporal Relation Encoding (TH-TRE) ◆ Temporal Relation-Aware Pooling (TRAP)	☐ **Experiments** ◆ SoTA Comparisons ◆ Ablation Experiments ◆ Real-time Performance ◆ More Experiments ◆ Conclusions

Contents.

I. Background --- Deep discussion and analysis of current temporal modeling paradigm

NEURAL INFORMATION PROCESSING SYSTEMS

Causal temporal relation modeling

Hidden Markov models (HMM) :

The classic causal temporal modeling methods (Fig. 1-a1):

- ✓ Introducing **hidden states** for temporal dependence modeling;
- ✓ Using **transition probabilities** between hidden states to describe the intrinsic causal relationship of sequential data;
- ➢ Fail to perform data representation and downstream task prediction **end-to-end**;
- ➢ The **limitation to describe long-term dependencies**.

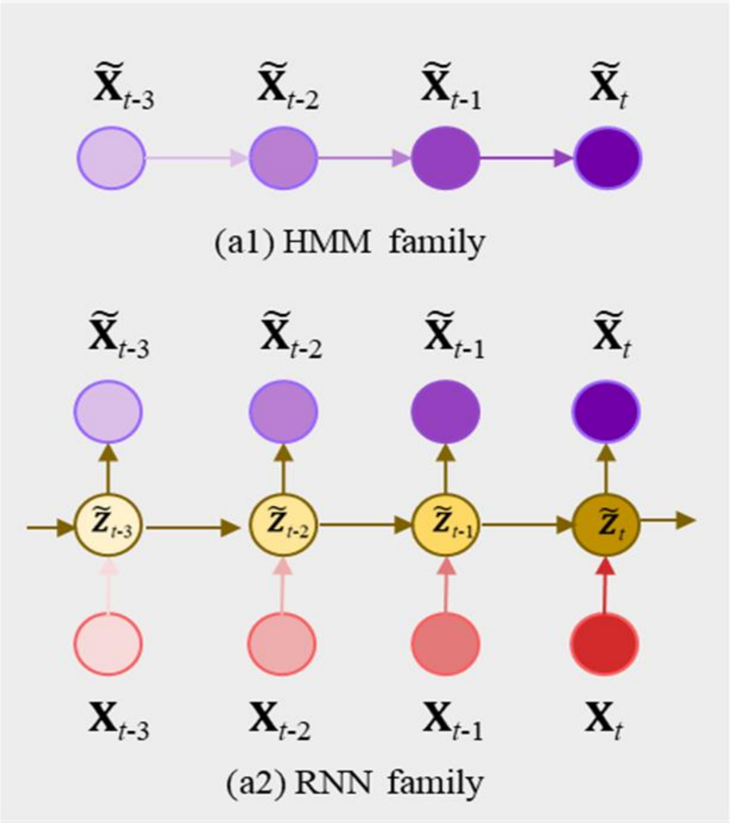

Fig.1 Causal temporal relation modeling

Recurrent neural network (RNN):

The deep learnable version of HMMs (Fig. 1-a2):

- ✓ Using **hidden state** to describe temporal relation of input sequence similar with HMMs;
- ✓ A **learnable model that can be deepened**;
- ➢ The resistance caused by **gradient dispersion;**
- ➢ The **limitation for paralleled computing**.

I. Background --- Deep discussion and analysis of current temporal modeling paradigm

NEURAL INFORMATION PROCESSING SYSTEMS

Parallelized sequence representation modeling

3D convolution(3DConv):

The popular spatio-temporal modeling component (Fig. 2-b1):

- ✓ **3D shaped local receptive field** (LRF) with shared kernel for representation of spatio-temporal tensors;
- ✓ The **fully paralleled computing** advantage inherit from convolution;
- ➢ The **trade-off** between long-term temporal-dependence and efficient computation with appropriate parameter amount.
- ➢ The rules of **context encoding** for 3DConv may not be optimal for RSS.

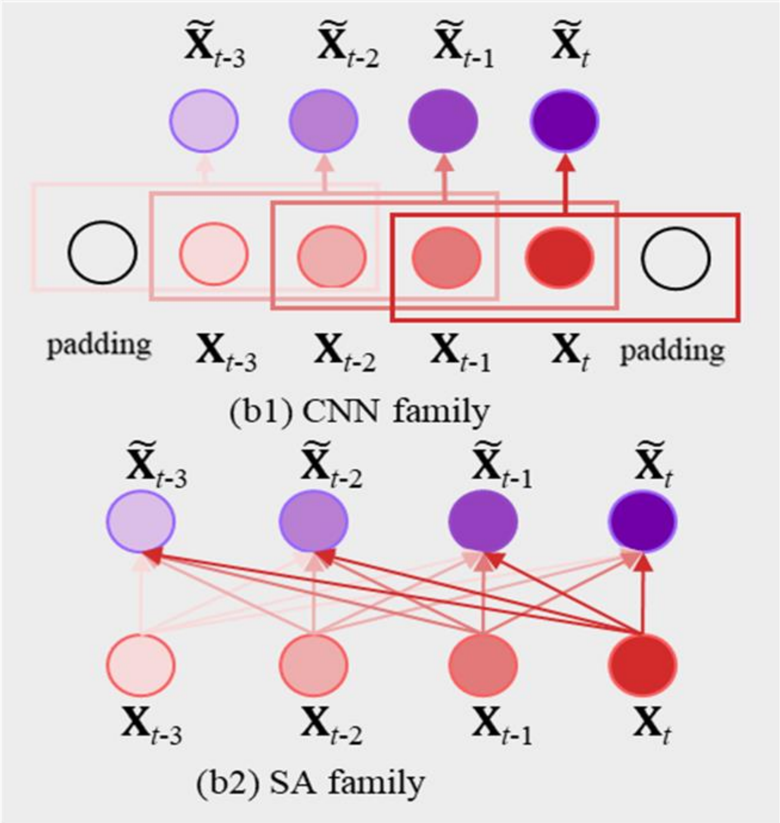

Fig.2 Parallelized sequence representation modeling

Transformer:

The backbone choice for modern fundamental models (Fig. 2-b2):

- ✓ Overcomes the problem of **parallel computing** for handling sequential data;
- ✓ **Breaks the limitation for long-term dependence** in LRF of 3DConv;
- ➢ The greedy relation computation introduces **high-cost for RSS** that requires the real-time performance for handeling high-dimensional spatio-temporal tensors.

Contents.

I. BACKGROUND	II. MOTIVATION	III. METHODOLOGY	VI. RESULTS
☐ **Deep discussion and analysis of current temporal modeling paradigm**	■ **Design principles of spatio-temporal encoding for Radar Semantic Segmentation**	☐ **Temporal Relation Attentive Model (TRAM)** ◆ Target-History Temporal Relation Encoding (TH-TRE) ◆ Temporal Relation-Aware Pooling (TRAP)	☐ **Experiments** ◆ SoTA Comparisons ◆ Ablation Experiments ◆ Real-time Performance ◆ More Experiments ◆ Conclusions

I. Motivation

Design principles of spatio-temporal encoding suitable for RSS domain

- **Utilizing parameterized rather than probabilistic connections to characterize the temporal relations** like Transformers and RNNs do.
- On the premise of making predictions at current time step, the module should **emphasize the use of current input frame**, i.e., non-context calculation in time.
- The module should be able to handle **temporal relations in parallel.**
- Considering the high-dimensional characteristics of radar data, the module should **ensure the efficient learning ability of long-term relationship and keep the parameters appropriately scaled.**
- Considering the non-smooth characteristics of radar data in time dimension, The module should **consider the contribution of each time step differently** during historical information aggregation.

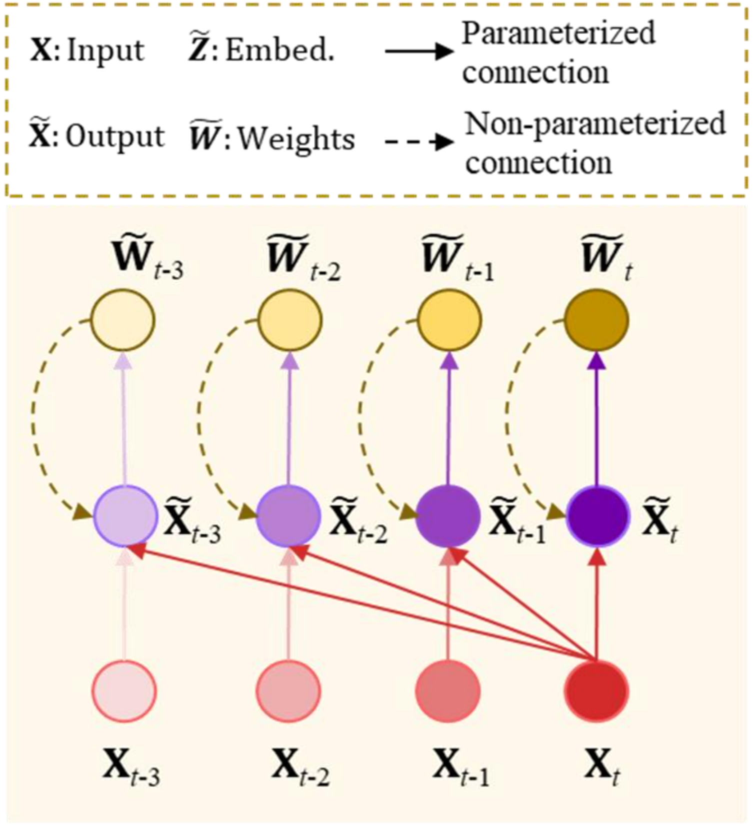

Fig.3 Proposed TRAM

Contents.

NEURAL INFORMATION PROCESSING SYSTEMS

III. Methodology

The proposed TARSS-Net is based on CAED framework, which consists of basic encoder, TRAM, latent space encoder (LSE) and decoder.

The LSE is used to align and fuse the high-level semantic features of different views, which is further applied to the single-view decoder to improve its performance.

Decoder receives inputs from TRAM and LSE, and finally produces segmentation results on RD and RA perspectives, respectively.

TRAM is the key of temporal relation learning.

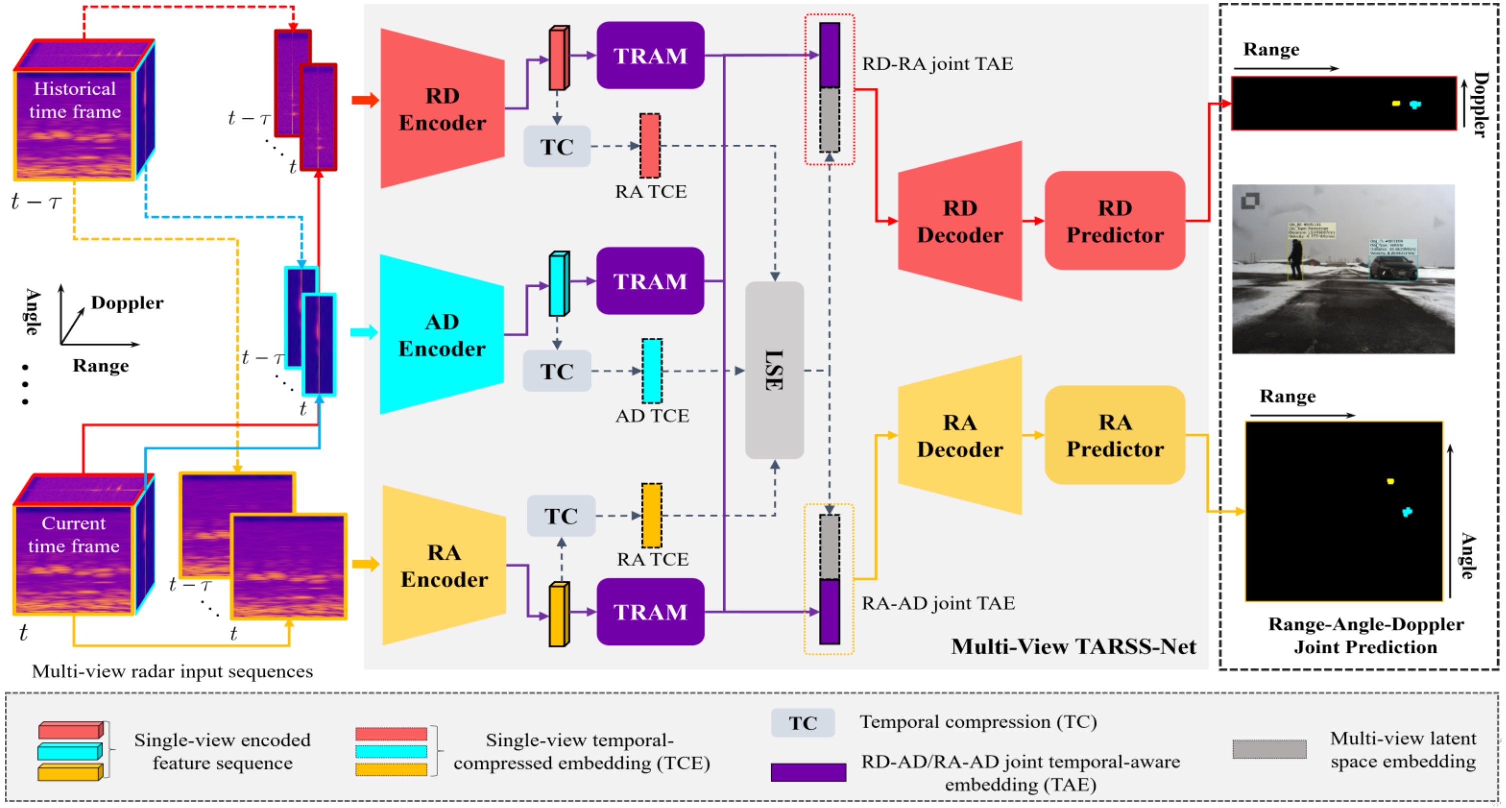

Fig.4 The overview of TARSS-Net

ISTA of CASIC
Intelligent Science & Technology

III. Methodology

Target-History Temporal Relation Encoding (TH-TRE)

- **TH-TRE aims at capturing temporal relations of encoded target frame and its adjacent historical frame features**. The designed temporal-relation-inception convolution (TRIC) block handles each target-historical feature pair, as shown in Fig.5.

Given the feature map sequence obtained from a basic encoder, the whole process of TH-TRE can be formalized as follow:

$$TH-TRE\left(\{X_j\}_{j=t-\tau}^{t}\right) = \{TRIC(X_t, \{X_j\}_{j=t-\tau}^{t-1}) \oplus^{\mathcal{T}} \mathbf{Max}(\mathcal{K}_1(\mathbf{X}_t))\},$$

$$\text{where } TRIC\left(X_t, \{X_j\}_{j=t-\tau}^{t-1}\right) = \{\mathcal{K}_2(\mathcal{K}_1(\mathbf{X}_t) \oplus^{\mathcal{D}} \mathcal{K}_1(\mathbf{X}_i))\}_{i=t-\tau}^{t-1}$$

Where, $\mathcal{K}_1$ and $\mathcal{K}_2$ are 2D convolution layers, $\oplus^{\mathcal{D}}$ and $\oplus^{\mathcal{T}}$ denotes concatenation on depth and temporal dimension, respectively, **Max** is the 2D max-pooling operation with the spatial downsampling rate of 2.

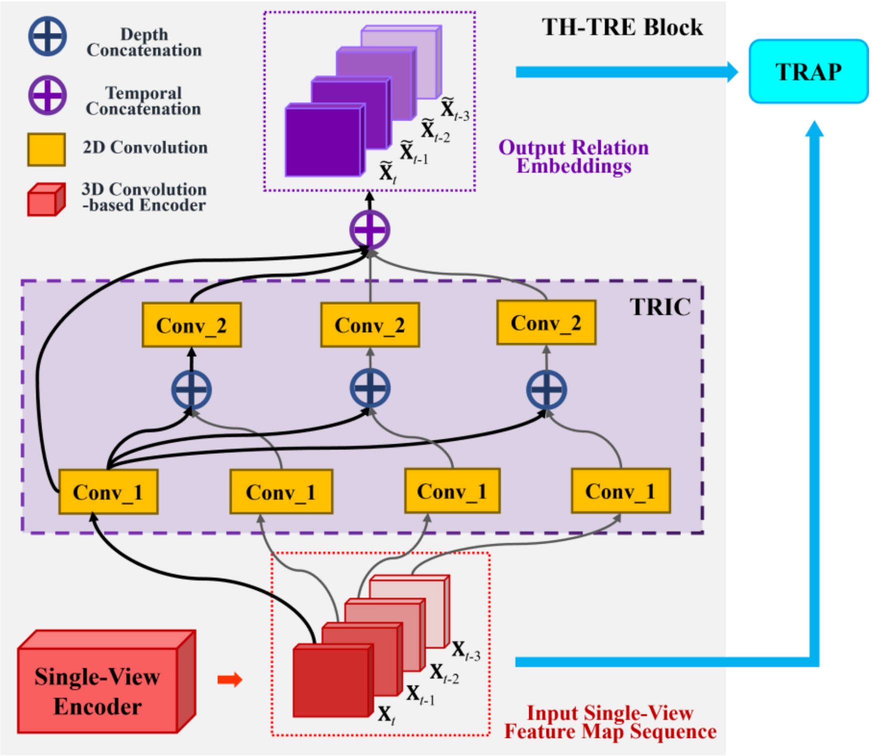

Fig.5 The illustration of TH-TRE

III. Methodology

Temporal Relation-Aware Pooling (TRAP)

- **TRAP block aims at perceiving the contribution degree of each historical frame for prediction task according to the target-history relations, and using these measurements of importance to aggregate the temporal information in each single-view radar sequence.** Two forms of TRAP are presented, i.e., Spatio-TRAP and Depth-TRAP.

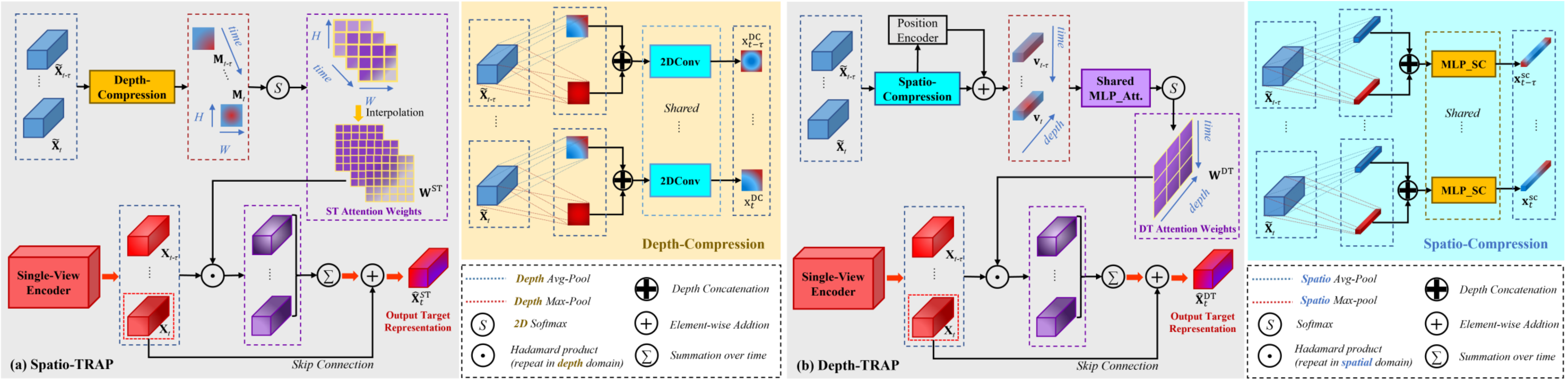

Fig.6 The illustration of two forms of TRAP block

III. Methodology

Temporal Relation-Aware Pooling (TRAP)

- **Spatio-TRAP** is performed on the entire spatial domain of input feature maps. Therefore, the importance of temporal relations will be estimated on the spatial space of relation embeddings.

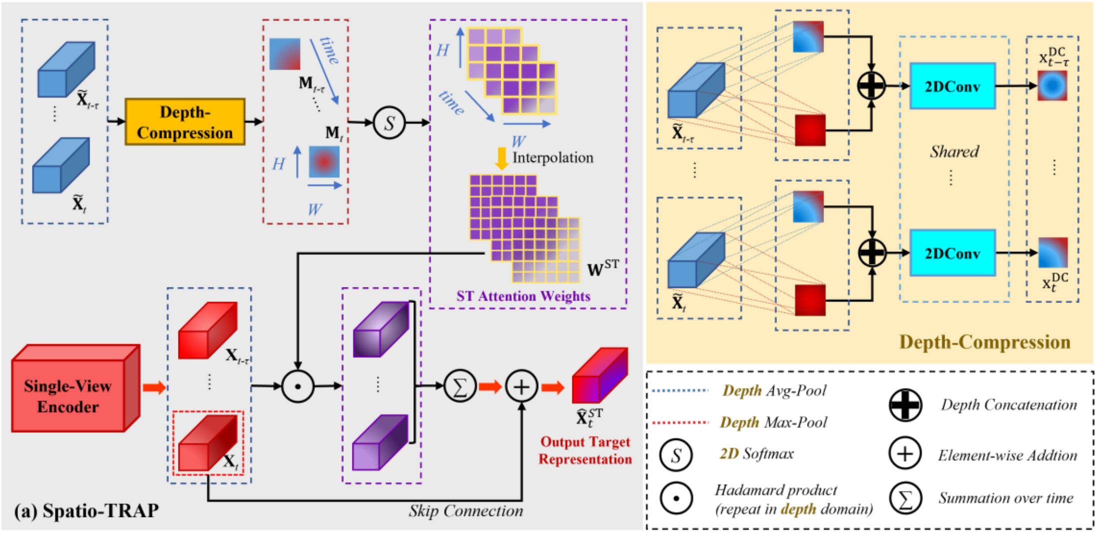

Fig.7 Spatio-TRAP

- **Depth-Compression:**

$$\mathbf{M}_i = \mathcal{K}_2^{ST}\left(\mathcal{K}_1^{ST}\left(\mathtt{Avg}^{DC}(\tilde{\mathbf{X}}_i) \oplus^{\mathcal{D}} \mathtt{Max}^{DC}(\tilde{\mathbf{X}}_i)\right)\right).$$

- **Spatio-Temporal Attentive Pooling:**

$$\hat{\mathbf{X}}_t^{ST} = \left\{\sum_{i=t-\tau}^{t} \mathbf{W}_i^{ST} \odot \mathbf{X}_{i,d}\right\}_{d=1}^{C} + \mathbf{X}_t, \text{ where,}$$

$$\mathbf{W}_i^{ST} = \frac{\mathtt{Intp}(\{\mathtt{2DSoftmax}(\mathbf{M})\}_i, [H/H_2, W/W_2])}{HW/H_2W_2}.$$

III. Methodology

Temporal Relation-Aware Pooling (TRAP)

- **Depth-TRAP** is performed on the depth of input feature maps. It measures the importance of temporal relations on semantic space of relation embeddings.

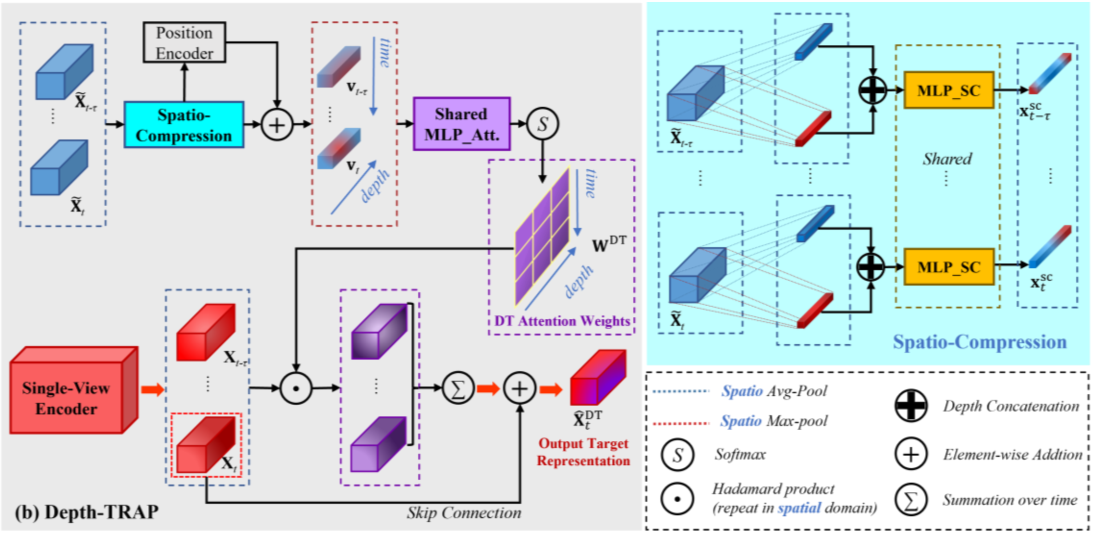

Fig.8 Depth-TRAP

- **Spatio-Compression:**

$$\mathbf{x}_i^{SC} = \mathcal{G}^{SC}\left(\mathtt{Avg}^{SC}(\tilde{\mathbf{X}}_i) \oplus^{\mathcal{D}} \mathtt{Max}^{SC}(\tilde{\mathbf{X}}_i)\right).$$

- **Depth-Temporal Attentive Pooling:**

$$\hat{\mathbf{X}}_t^{DT} = \left\{\sum_{i=t-\tau}^{t} \mathbf{W}_i^{DT} \odot \mathbf{x}_{i,h,w}\right\}_{h=1,w=1}^{H,W} + \mathbf{X}_t.$$

$$\mathbf{W}^{DT} = \mathtt{Softmax}\left(\{\mathcal{G}^{DT}(\mathbf{v}_i)\}_{i=t-\tau}^{t}\right) \text{ where,}$$

$$\mathbf{v}_i = \{x_{i,d}^{SC} + p_{i,d}\}_{d=1}^{C},\; p_{i,d} = \begin{cases} p_{i,d} = 0.1\sin\left(i/10^{8(d/2)/C}\right), & \text{if } d \bmod 2 = 0; \\ p_{i,d} = 0.1\cos\left(i/10^{8((d-1)/2)/C}\right), & \text{if } d \bmod 2 = 1. \end{cases}$$

Contents.

IV. Experiments & Results

- The effectiveness of TARSS-Nets compared with the existing state-of-the-art algorithms is verified on different datasets and it has outstanding performance on all three test datasets.
- With the careful design of temporal modeling paradigm, TARSS-Net can handle arbitrarily long sequence relations without increasing parameter scale. Effect of input time length on TARSS-Net performance is shown in Fig.9.

Tab.1 Comparisons with SoTA RSS networks.

Method	#Param.	RD-View (%)		RA-View (%)	
		mIoU	mDice	mIoU	mDice
FCN[20]	134.3M	54.7	66.3	34.5	40.9
U-Net[28]	17.3M	55.4	68.0	32.8	38.2
DeepLabv3+[2]	59.3M	50.8	61.6	32.7	38.3
RSS-Net[14]	10.1 M	32.1	36.9	32.1	37.8
RAMP-CNN[6]	106.4 M	56.6	68.5	27.9	30.5
MV-Net[21]	**2.4 M**	29.0	32.8	26.8	28.5
MVA-Net[21]	4.8 M	53.5	65.3	37.1	44.8
TMVA-Net[21]	5.6 M	56.1	68.0	37.7	46.2
TransRadar[5]	4.9M	57.2	69.1	39.9	49.5
T-RODNet[13]	162.0M	-	-	**43.5**	53.6
TransRSS[36]	-	60.4	73.0	43.0	**53.8**
PKCIn-Net[35]	6.3M	60.7	72.6	43.1	53.7
TARSS-Net_S	6.2 M	62.1	73.8	41.6	51.2
TARSS-Net_D	6.3 M	**63.4**	**75.2**	41.4	51.3

Tab.2 Performance on CARRADA-RAC.

View	Method	#Param.	mIoU	mDice
RD	TMVA-Net	**5.6M**	59.7%	69.9%
	PKCIn-Net	6.3M	60.6%	72.4%
	TARSS-Net_S	6.2M	62.5%	74.3%
	TARSS-Net_D	6.3M	**62.8%**	**74.6%**
RA	TMVA-Net	**5.6M**	46.6%	57.9%
	PKCIn-Net	6.3M	47.3%	**58.7%**
	TARSS-Net_S	6.2M	45.8%	56.1%
	TARSS-Net_D	6.3M	**47.4%**	**58.7%**
Global	TMVA-Net	**5.6M**	53.2%	63.9%
	PKCIn-Net	6.3M	54.0%	65.6%
	TARSS-Net_S	6.2M	54.1%	65.2%
	TARSS-Net_D	6.3M	**55.1%**	**66.7%**

Tab.3 Performance on KuRALS.

Method	#Param.	RD View	
		mIoU	mDice
FCN	134.3M	50.4%	59.4%
U-Net	17.3M	52.4%	60.1%
DeepLabv3+	59.3M	52.6%	61.8%
TMVA-Net^{sv}	**1.2M**	52.9%	63.1%
PKCIn-Net^{sv}	**1.2M**	56.7%	65.9%
TARSS-Net_S^{sv}	**1.2M**	53.2%	63.8%
TARSS-Net_D^{sv}	1.3M	**58.4%**	**67.1%**

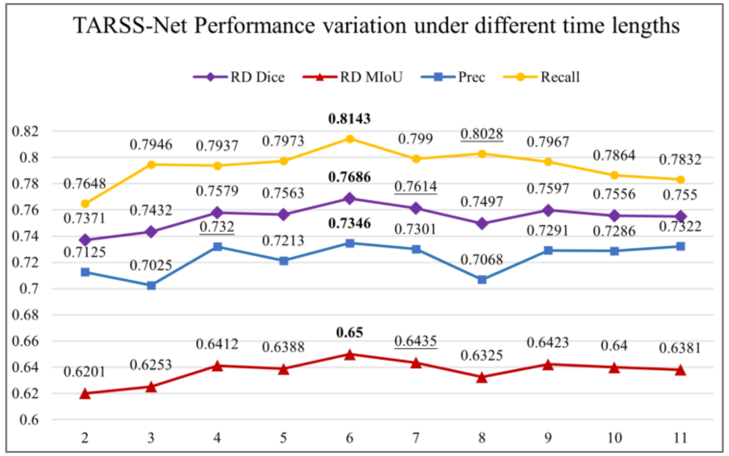

Fig.9 TARSS-Net performance using input sequence with different numbers of frames.

NEURAL INFORMATION PROCESSING SYSTEMS

IV. Experiments & Results

- Ablation Study
  - Ablation Experiments on TRAM
  - The Effectiveness of TH-TRE
  - The Effectiveness of TRAP

- Real-time performance

  It is measured by multiply–accumulate operations (MACs) and frames per second (FPS), respectively. All the real-time performance shown in this section are obtained on a single RTX 3090 GPU.

Tab.4 Ablation experimental results on TRAM.

Method	RD-View (%)				RA-View (%)				Global (%)			
	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice
Baseline-A w/ GAP	61.8	76.3	55.4	67.8	43.5	47.1	36.4	44.7	52.7	61.7	45.9	56.3
Baseline-A w/ GMP	68.6	68.2	50.8	62.1	49.3	42.9	36.1	42.7	59.0	55.6	43.5	52.4
Baseline-B	63.6	74.8	56.1	68.0	44.2	51.6	37.7	46.2	53.9	63.2	46.9	57.1
TARSS-Net w/ TRAM	70.9	80.6	63.4	75.2	56.1	50.0	41.4	51.3	63.5	65.3	52.4	63.3

Tab.5 Ablation experimental results on TH-TRE.

Method	RD-View (%)				RA-View (%)				Global (%)			
	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice
Baseline-A w/ GAP	61.8	76.3	55.4	67.8	43.5	47.1	36.4	44.7	52.7	61.7	45.9	56.3
Baseline-A w/ GAP & TH-TRE	68.9	79.3	60.6	72.3	50.9	50.2	40.4	49.7	59.9	64.8	50.5	61.0
TARSS-Net w/o TH-TRE	69.9	79.2	61.6	73.7	53.2	50.1	40.4	49.8	61.6	64.7	51.0	61.8
TARSS-Net w/ TH-TRE	70.9	80.6	63.4	75.2	56.1	50.0	41.4	51.3	63.5	65.3	52.4	63.3

Tab.6 The effects of Depth/Spatio-TRAP block.

Method	RD-View (%)				RA-View (%)				Global-View (%)			
	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice	Prec.	Recall	mIoU	mDice
Baseline-A_G	61.8	76.3	55.4	67.8	43.5	47.1	36.4	44.7	52.7	61.7	45.9	56.3
Baseline-A_S	69.5	78.5	61.0	73.3	50.8	49.4	38.8	47.4	60.2	64.0	49.9	60.4
Baseline-A_D	69.9	79.2	61.6	73.7	53.2	50.1	40.4	49.8	61.6	64.7	51.0	61.8
TARSS-Net_G	68.9	79.3	60.6	72.3	50.9	50.2	40.4	49.7	59.9	64.8	50.5	61.0
TARSS-Net_S	70.4	79.1	62.1	73.8	53.6	50.6	41.6	51.2	62.0	64.9	51.9	62.5
TARSS-Net_D	70.9	80.6	63.4	75.2	56.1	50.0	41.4	51.3	63.5	65.3	52.4	63.3

Tab.7 Real-time performance (MV: multi-view; SV: single-view).

Method	Inputs	Params (M)	MACs (G)	FPS	mIoU (%)	mDice (%)	Inputs	Params (M)	RD-View		RA-View	
									MACs(G)	FPS	MACs(G)	FPS
TMVA-Net	MV	7.2	119.5	66	46.9	57.1	SV	1.2	11.3	250	36.6	250
Vit-based-Net	MV	27	449	12	38.1	44.5	SV	3.6	1.8	59	7.0	55
TARSS-Net_S	MV	6.2	197.6	35	51.9	62.5	SV	1.2	13.3	181	40.4	143
TARSS-Net_D	MV	6.3	175.4	23	52.4	63.3	SV	1.2	13.3	111	40.4	112

ISTA of CASIC
Intelligent Science & Technology

IV. Experiments & Results

Visualization results

- Feature Visualization
- Visualization of some examples

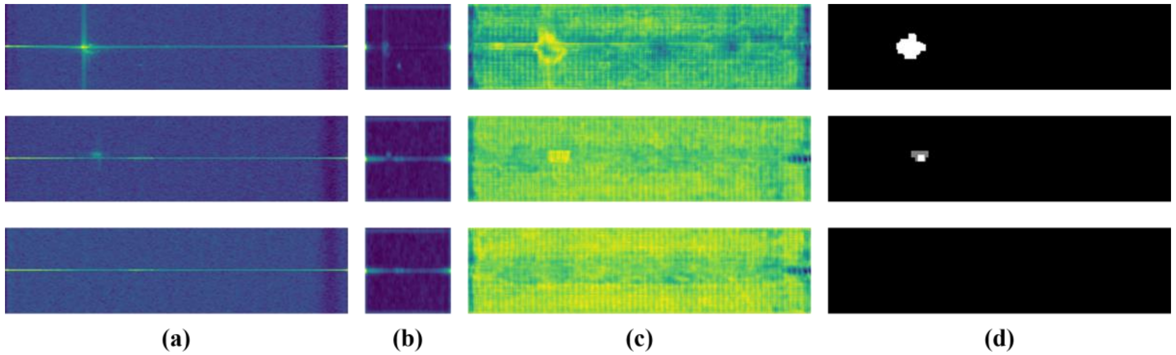

Fig.10 Feature Visualization. (a) Input RD-view frame. (b) The activation response heatmaps of TRAM outputs. (c) TARSS-Net outputs before Softmax. (d) Ground Truth Mask.

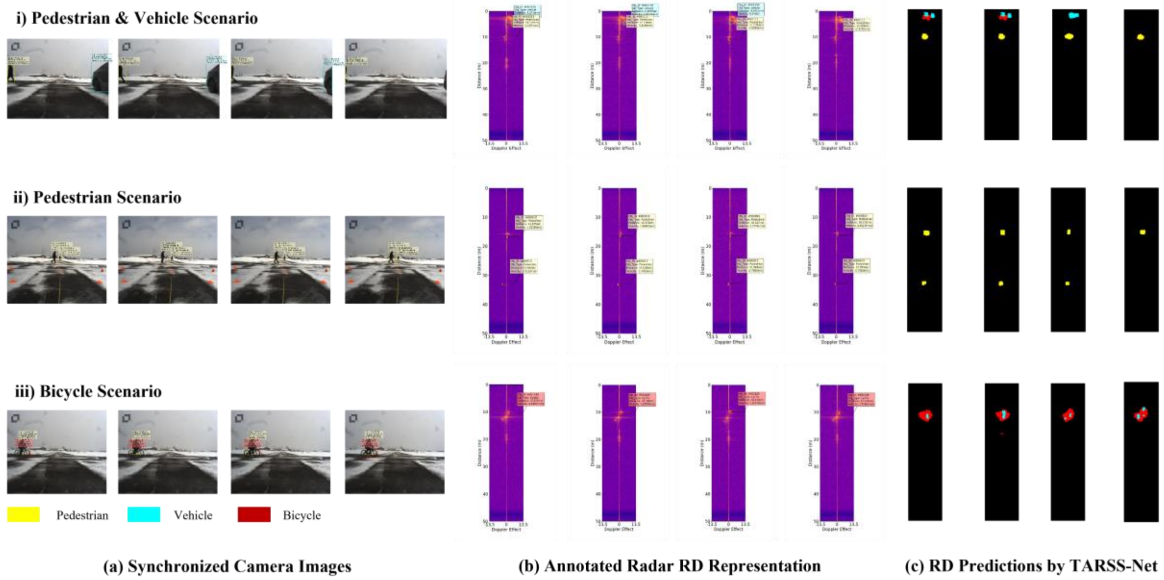

Fig.11 Visualization of some examples

NEURAL INFORMATION PROCESSING SYSTEMS

Conclusions

TARSS-Net focuses on exploiting temporal information in radar signals to enhance the representation capacity of RSS model:

I. The existing temporal modeling methods in RSS were deeply discussed;
II. The design principles of RSS spatio-temporal encoding methods were introduced;
III. A flexible temporal-aware learning module, TRAM, and TARSS-Net based on TRAM is proposed, following the proposed temporal learning paradigm, i.e., **data-driven temporal information aggregation with learned target-history relations;**
IV. Experiments fully verifies the superiority of TARSS-Net through SoTA methods comparison on three datasets, ablation experiments, performance under variation input time length, as well as its real-time performance.

NEURAL INFORMATION PROCESSING SYSTEMS

ISTA of CASIC
Intelligent Science & Technology

Thanks for your attention!

Q&A

YOUCHENG ZHANG*, LIWEN ZHANG*,
TENG LI, ET AL.
DEC. 2024.

Question and cooperation please connect with lwzhang9161@126.com
TARSS-Net Project: https://github.com/zlw9161/TARSS-Net