Deformba: Vision State Space Model with Adaptive State Fusion

Hongyu Ke[1] Jack Morris[2] Yongkang Liu[3] Satoshi Kitai[3] Kentaro Oguchi[3] Yi Ding[2] Haoxin Wang[1]

Georgia State University[1] University of Tennessee, Knoxville[2] InfoTech Labs, Toyota Motor North America R&D[3]

Overview

- **What?** Introduce context-adaptive state fusion to enable Vision SSMs dynamically read relevant spatial and query-source evidence from a shared state memory.
- **Why?** Fixed-scan Vision SSMs impose rigid 1D ordering on 2D images and lack natural query-based fusion, limiting spatial modeling and multi-view perception.
- **Where?** Our code is open source! We hope to provide a strong baseline and evaluation framework for future experimentation.

Method: Write Once, Read Many

Deformba decouples SSM computation into a **write pass** and an adaptive **read pass**. Given visual features $F \in \mathbb{R}^{H\times W\times C}$, a single SSM scan writes them into a shared 2D state memory:

$$S_t = \alpha_t S_{t-1} + v_t k_t^\top, \qquad S_{2D} \in \mathbb{R}^{H\times W\times C}.$$

Then, CASF adaptively reads relevant evidence from the state map. An offset prediction network first estimates where each token should sample from the shared memory:

$$\Delta p = f_{\text{off}}(S_{2D}).$$

The sampling locations are obtained by adding the predicted offsets to the reference points:

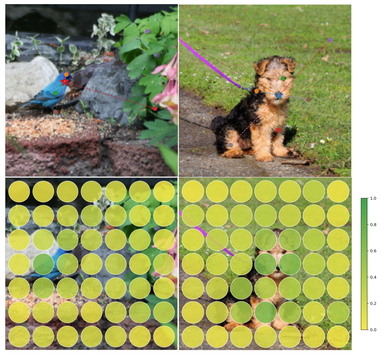
$$P = E + \Delta p.$$

The sampled state features are aggregated with learned weights:

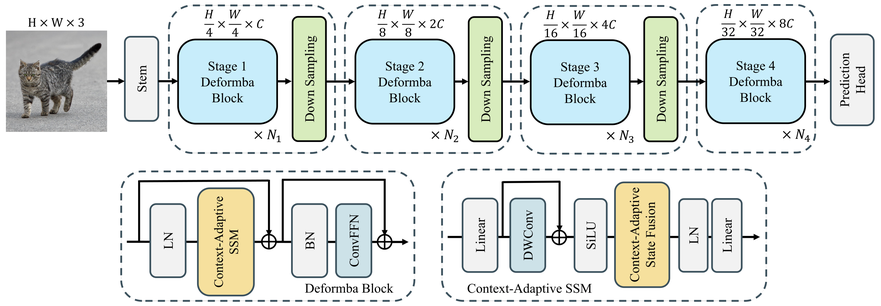
$$S_Q = \sum_{g=1}^{G} w_g \cdot \phi(S_{2D}, P_g).$$

Finally, the enriched state is projected to generate the output:

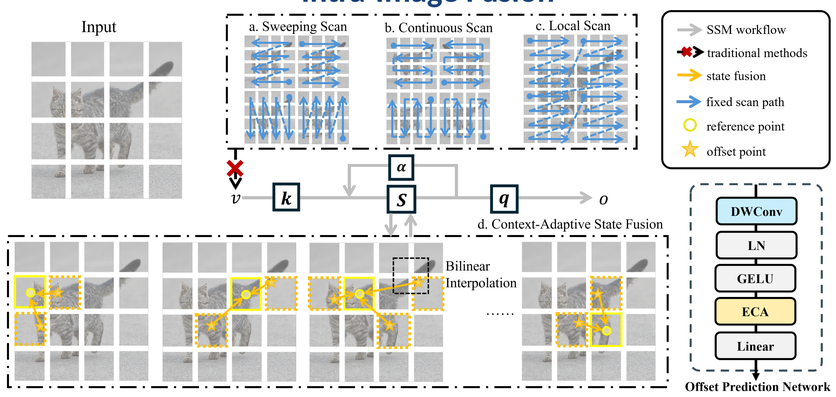
$$O = S_Q Q + V D.$$



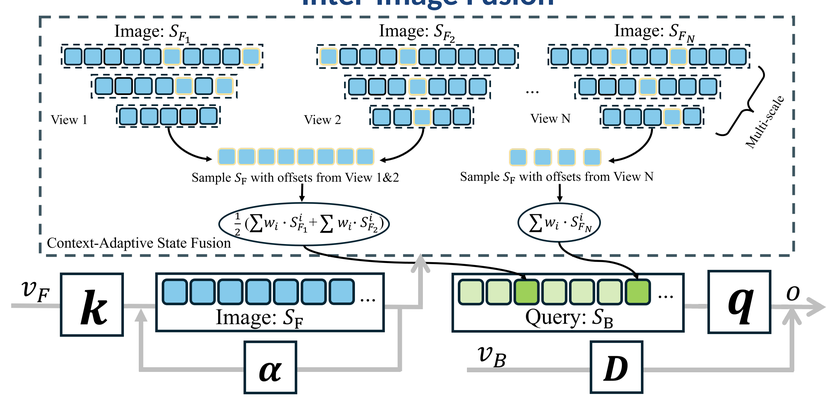
Architecture



Intra-Image Fusion



Inter-Image Fusion



Key Points

- **Adaptive Read:** Learns where to gather spatial evidence instead of relying on fixed scan orders.
- **Efficient SSM Fusion:** Preserves linear SSM efficiency through a write-once, adaptive-read design.
- **General Vision Framework:** Supports both 2D perception and cross-attention-style fusion for 3D BEV tasks.

Experiment Results

Classification results on ImageNet Dataset

Method	Type	Params(M)↓	FLOPs(G)↓	Top-1(%)↑
UniRepLKNet-T	CNNs	31	4.9	83.2
InternImage-T	CNNs	30	5.0	83.5
Swin-T	ViTs	28	4.5	81.3
CMT-S	ViTs	25	4.0	83.5
VMamba-T	SSMs	22	5.6	82.6
DefMamba-S	SSMs	32	4.8	83.5
Deformba-T	SSMs	25	4.8	**83.8**
UniRepLKNet-S	CNNs	56	9.1	83.9
InternImage-S	CNNs	50	8.0	84.2
Swin-S	ViTs	50	8.7	83.0
TransNeXt-S	ViTs	50	10.3	84.7
VMamba-S	SSMs	44	11.2	83.6
DefMamba-B	SSMs	51	8.5	84.2
Deformba-S	SSMs	45	10.3	**84.9**
MambaOut-B	CNNs	85	15.8	84.2
InternImage-B	CNNs	97	16.0	84.9
SwinV2-B	ViTs	88	15.1	84.6
TransNeXt-B	ViTs	90	18.4	84.8
VMamba-B	SSMs	75	18.0	83.9
Deformba-B	SSMs	85	16.3	**85.4**

BEV 3D Detection results on nuScenes Dataset

Method	Backbone	# Frames	NDS↑	mAP↑
BEVFormerV1-Tiny	ResNet50	3	0.354	0.252
BEVFormerV2-Tiny	ResNet50	3	0.397	0.270
MamBEV-Tiny	ResNet50	3	0.399	0.266
Deformba-Tiny	ResNet50	3	**0.405**	**0.276**
PolarDETR-T	ResNet101	2	0.488	0.383
BEVFormerV1-Small	ResNet101	3	0.479	0.370
BEVFormerV1-Base	ResNet101	4	0.517	0.416
MamBEV-Small	ResNet101	3	0.523	0.415
Deformba-Small	ResNet101	3	**0.525**	**0.432**