ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer

Haoran You*, Huihong Shi*, Yipin Guo*, Yingyan (Celine) Lin

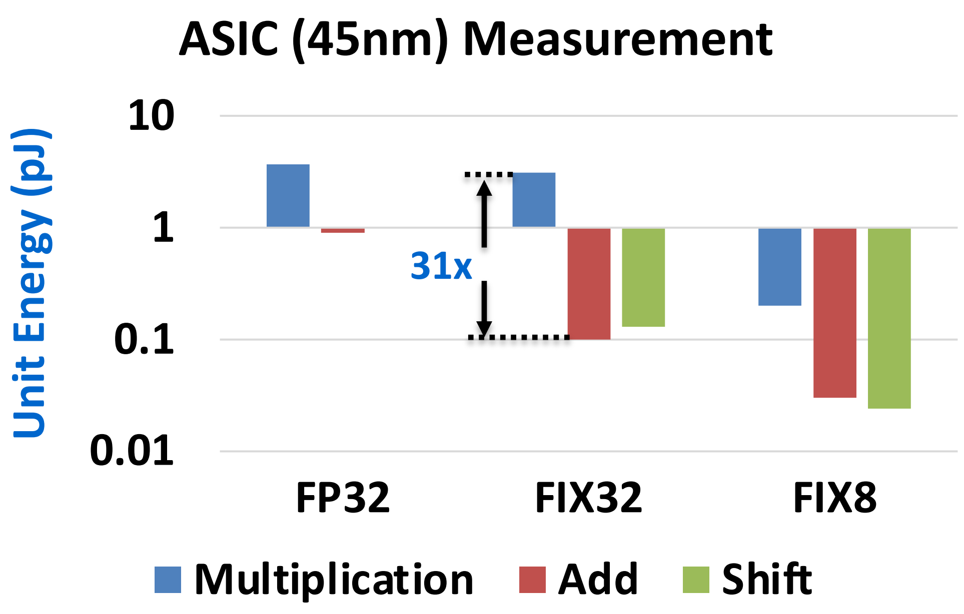
NeurIPS 2023

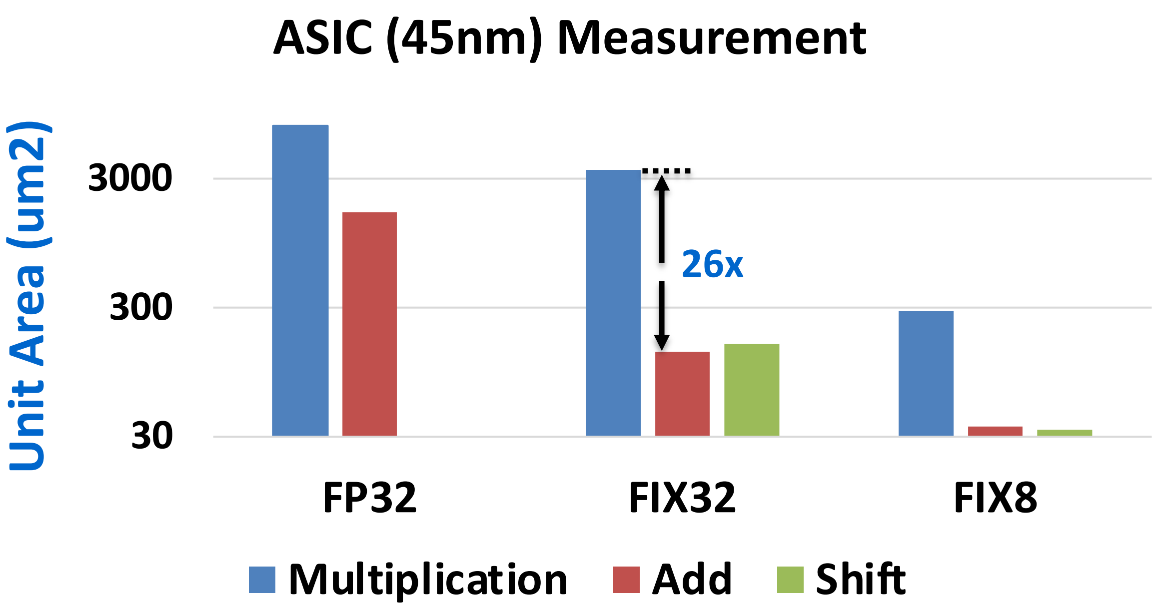
Georgia Institute of Technology

Efficient and Intelligent Computing Lab

ShiftAddViT: Background and Motivation

- **Powerful Vision Transformers (ViTs) suffer from large inference and training cost**
  - **Bottleneck:**
    - Both attentions and MLPs are not efficient enough due to dense multiplications
  - **Opportunity:**
    - Multiplication can be represented by bitwise shift and add
      (e.g., up to **31x unit energy reduction** **and** **26x unit area reduction** over multiplication)





ShiftAddViT: Background and Motivation

- **Powerful Vision Transformers (ViTs) suffer from large inference and training cost**
  - **Motivation:**
    - Reparametrize the pre-trained ViTs with mixture of multiplication primitives
  - **Prior Work:**

AdderNet *[H. Chen, CVPR'19]*, **DeepShift** *[M. Elhoushi, CVPRW'21]*, **Adder Attention** *[H. Shu, NeurIPS'22]*, **etc**

For CNNs or Transformers

Characteristics	Drawbacks
Less expressive capacity	↓ Accuracy
Training from scratch	↑ Training costs
Slow training & inference speed on GPUs	↓ Practical Usage

ShiftAddNet
[H. You, NeurIPS'20]

Dedicated for CNNs

Characteristics	Drawbacks
Cascaded shift & add layers	↑ Params & latency
Training from scratch	↑ Training costs
Slow training & inference speed on GPUs	↓ Practical Usage

ShiftAddViT: Background and Motivation

- **Powerful Vision Transformers (ViTs) suffer from large inference and training cost**
  - **Motivation:**
    - Reparametrize the pre-trained ViTs with mixture of multiplication primitives

ShiftAddViT
The Proposed Method

Dedicated for ViTs

Characteristics	Advantages
High expressive capacity	↑ Accuracy
Keep same number of layers	↓ Params & latency
Inherit pre-trained weights	↓ Training costs
Provide GPU optimizations	↑ Practical Usage

ShiftAddViT: Background and Motivation

- **Powerful Vision Transformers (ViTs) suffer from large inference and training cost**
  - **Motivation:**
    - Reparametrize the pre-trained ViT with mixture of multiplication primitives
  - **Associated Challenges:**
    - How to effectively reparameterize ViTs with shifts and adds?
    - How to maintain the accuracy after reparameterization?

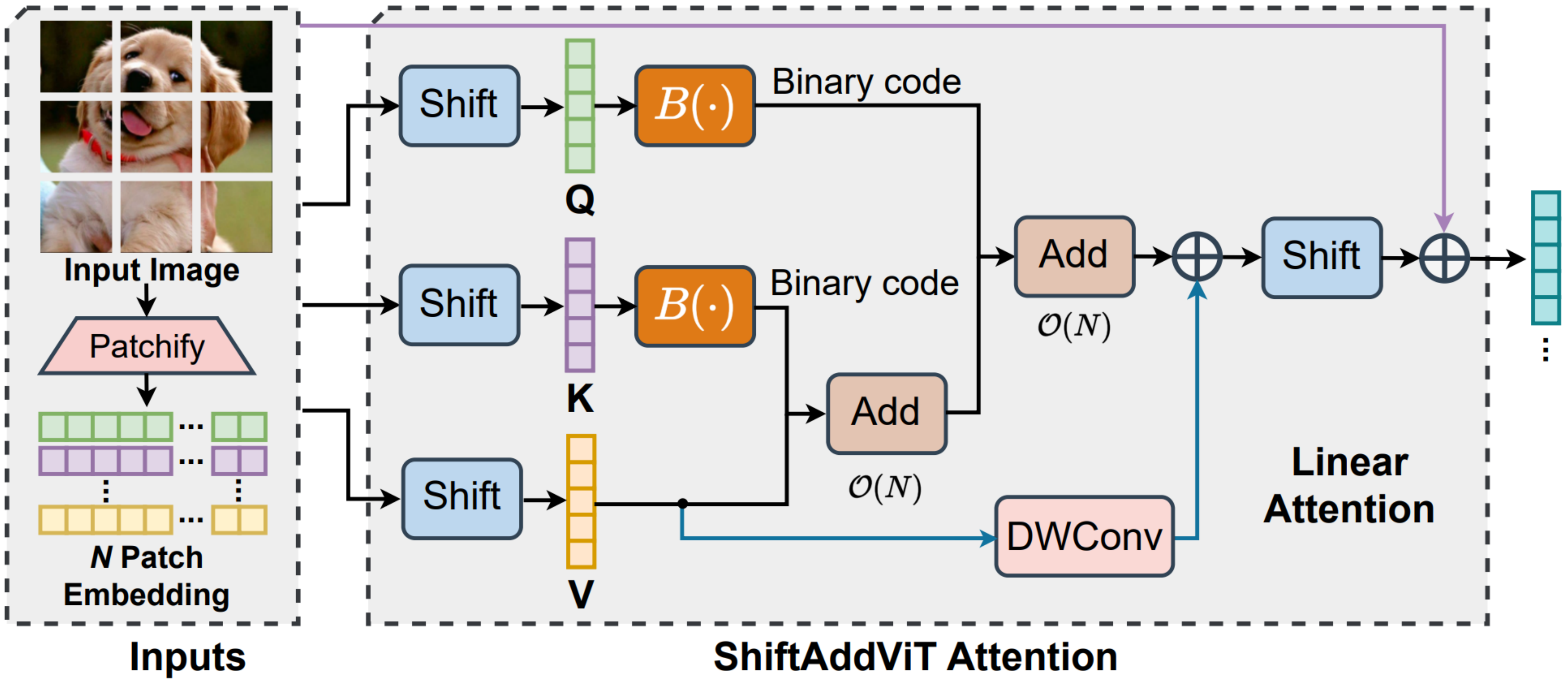
ShiftAddViT: Our Contributions

For the first time, we

- Reparameterize pre-trained ViTs with shifts and adds to deliver a new type of multiplication-reduced network, called ShiftAddViT
- Propose a new mixture of experts (MoE) framework for ShiftAddViT to preserve accuracy after reparameterization
- Introduce a latency-aware load-balancing loss term within our MoE framework to dynamically allocate input tokens to each expert
- Conduct extensive experiments on 2D and 3D vision tasks to validate the effectiveness of our proposed ShiftAddViT

Contribution 1: Reparameterization of Pre-trained ViTs

- **ShiftAddViT**
  - Reparameterization of Attention



Contribution 1: Reparameterization of Pre-trained ViTs

- **ShiftAddViT**
  - **Reparameterization of Attention**
  - **Reparameterization of MLPs**

Can we reparameterize all MLPs with Shifts?

Contribution 1: Reparameterization of Pre-trained ViTs

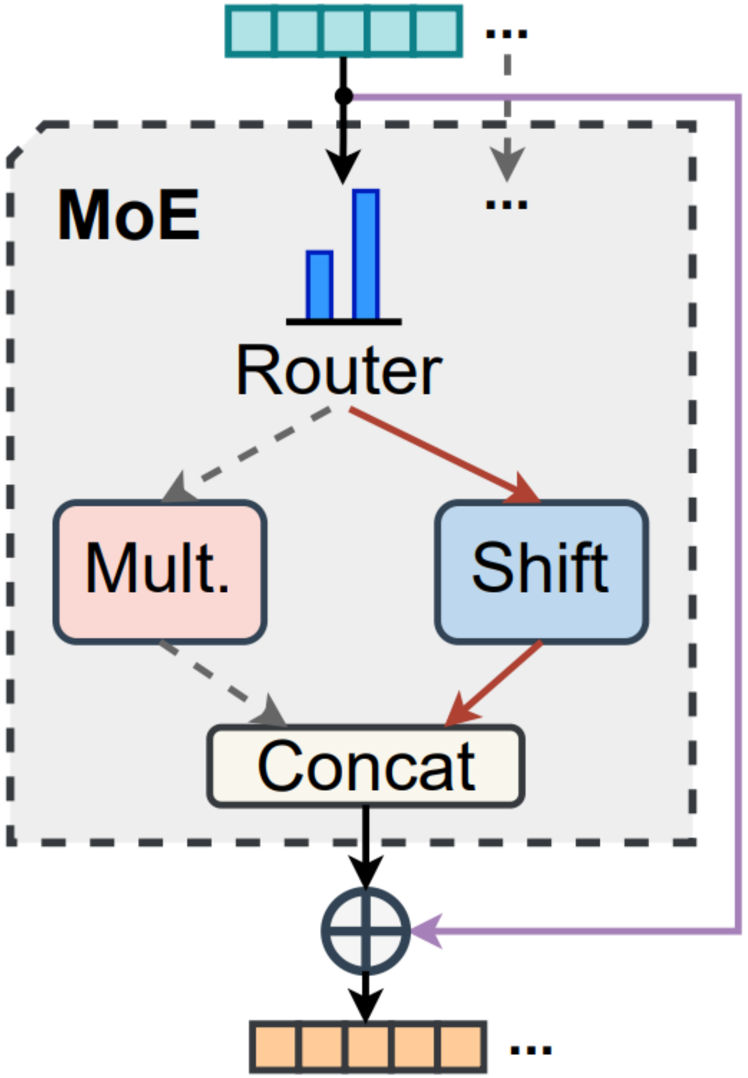
- **ShiftAddViT**
  - **Reparameterization of Attention**
  - **Reparameterization of MLPs**
    - **Sensitivity analysis**

Components	Apply	PVTv2-B0	PVTv1-T
-	MSA	71.25	76.21
Attention	Shift & Add	70.96	76.05
MLPs	Shift	70.28	73.92

Contribution 2: Mixture of Experts Framework

- **ShiftAddViT**
  - **Reparameterization of Attention**
  - **Reparameterization of MLPs**
    - **Hypothesis**

      Sensitive tokens need higher precision
      Insensitive tokens can be handled by cheaper primitives



ShiftAddViT MLP

Contribution 2: Mixture of Experts Framework

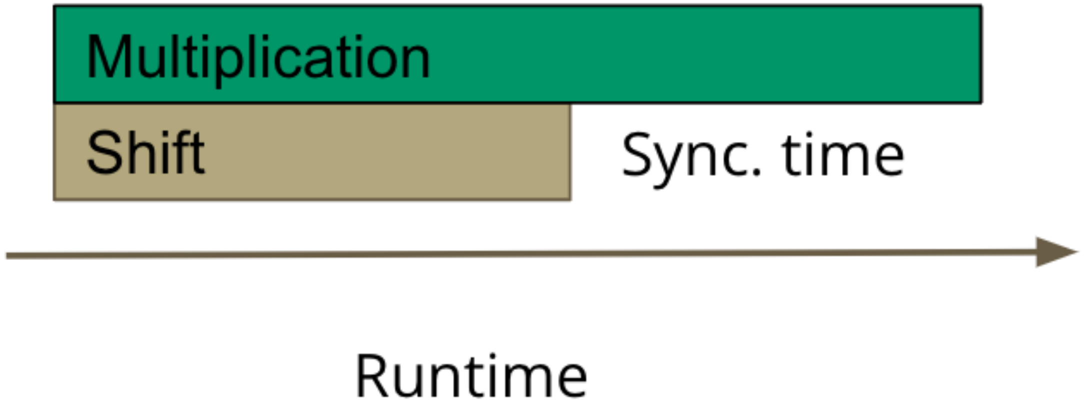
- **ShiftAddViT**
  - **Reparameterization of Attention**
  - **Reparameterization of MLPs**
    - **Sensitivity analysis**

Components	Apply	PVTv2-B0	PVTv1-T
-	MSA	71.25	76.21
Attention	Shift & Add	70.96	76.05
MLPs	Shift	70.28	73.92
	MoE	70.86	74.81

Contribution 3: Latency-aware Load-balancing Loss

- **ShiftAddViT**
  - Reparameterization of Attention
  - Reparameterization of MLPs

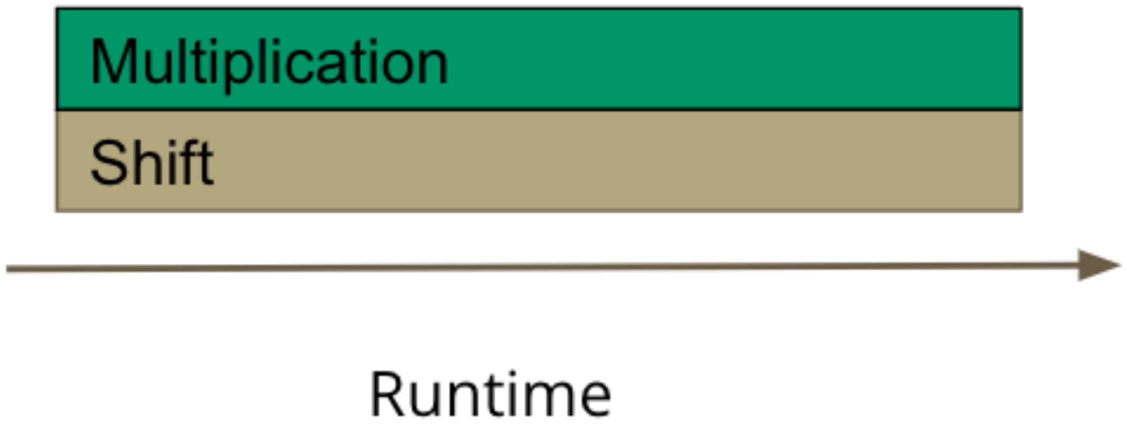
How to reduce the synchronization time in ShiftAddViT MoE?



Contribution 3: Latency-aware Load-balancing Loss

- **ShiftAddViT**
  - Reparameterization of Attention
  - Reparameterization of MLPs
  - Latency-aware Load-balancing Loss

Idea: The number of tokens assigned to experts aligns with the processing speeds of the experts



Experiment Setup and Baselines

- **Evaluation Setup**
  - **Tasks**: 2D Image Classification and 3D Novel View Synthesis
  - **Datasets**: ImageNet and Local Light Field Fusion (LLFF) with 8 scenes
  - **Models**: PVTv1, PVTv2, DeiT, and GNT
  - **Metrics**: Accuracy, Latency, and Energy

- **Benchmark Baselines**
  - **2D Transformers**
    - **Ecoformer** [J. Liu et al, NeurIPS'22]
    - **PVT** [W. Wang, ICCV'21]
  - **3D Transformers**
    - **NeRF** [B. Mildenhall, ECCV'22]
    - **GNT** [M. Varma, NeurIPS'22]

ShiftAddViT: Experimental Results for 2D Tasks

Models	**Methods**	**Acc. (%)**	**Latency (ms)**	**Energy (mJ)**
PVTv2-B0	Ecoformer [34]	70.44	7.82	33.64
	ShiftAddViT	**70.59**	**1.51**	**27.13**
PVTv1-T	Ecoformer [34]	NaN	7.43	93.47
	ShiftAddViT	**74.93**	**1.97**	**72.59**
PVTv2-B1	Ecoformer [34]	78.38	8.02	106.2
	ShiftAddViT	**78.49**	**2.49**	**85.34**
PVTv2-B2	Ecoformer [34]	81.28	15.43	198.2
	ShiftAddViT	**81.32**	**4.83**	**163.9**
DeiT-T	MSA [55]	72.20	5.12	66.88
	ShiftAddViT	**72.40**	**2.94**	**38.21**

Methods	**Linear Attn**	**Add**		**Shift**	**MoE**	**PVTv2-B0 [61]**			**PVTv1-T [60]**		
		KSH	**Quant.**			**Acc. (%)**	**Lat. (ms)**	**T. (img./s)**	**Acc. (%)**	**Lat. (ms)**	**T. (img./s)**
MSA	✗	✗	✗	✗	✗	70.77	4.62	989	76.21	4.73	903
PVT [61]	✓	✗	✗	✗	✗	70.50	6.25	2227	75.10	5.78	1839
PVT+MoE	✓	✗	✗	✗	✓ (MLPs)	70.82	12.46	1171	75.27	10.91	834
Ecoformer [34]	✓	✓	✗	✗	✗	70.44	7.82	1348	NaN	7.43	1021
ShiftAddViT (with KSH [34] or Quant. [27] to binarize Q/K)	✓	✗	✗	✗	✗	71.19	6.13	2066	75.50	5.78	1640
	✓	✓	✗	✗	✗	70.95	1.07†	2530†	75.20	1.42†	1683†
	✓	✓	✗	✓ (Attn)	✗	70.53	1.04†	2447†	74.77	1.39†	1647†
	✓	✓	✗	✓ (Attn)	✓ (MLPs)	70.16	1.39†/1.11*	N/A	74.44	1.91†/1.21*	N/A
	✓	✓	✗	✗	✓ (Both)	70.38	1.59†/1.20*	N/A	74.73	2.12†/1.21*	N/A
	✓	✗	✗	✗	✗	71.36	6.34	2014	75.64	5.48	1714
	✓	✗	✓	✗	✗	71.04	1.00†	2613†	75.18	1.20†	1907†
	✓	✗	✓	✓ (Both)	✗	68.57	0.97†	2736†	73.47	1.18†	1820†
	✓	✗	✓	✗	✓ (Both)	70.59	1.51†/1.12*	N/A	74.93	1.97†/1.02*	N/A

* denotes the modularized latency simulated by separately optimizing each expert/router with ideal parallelism.

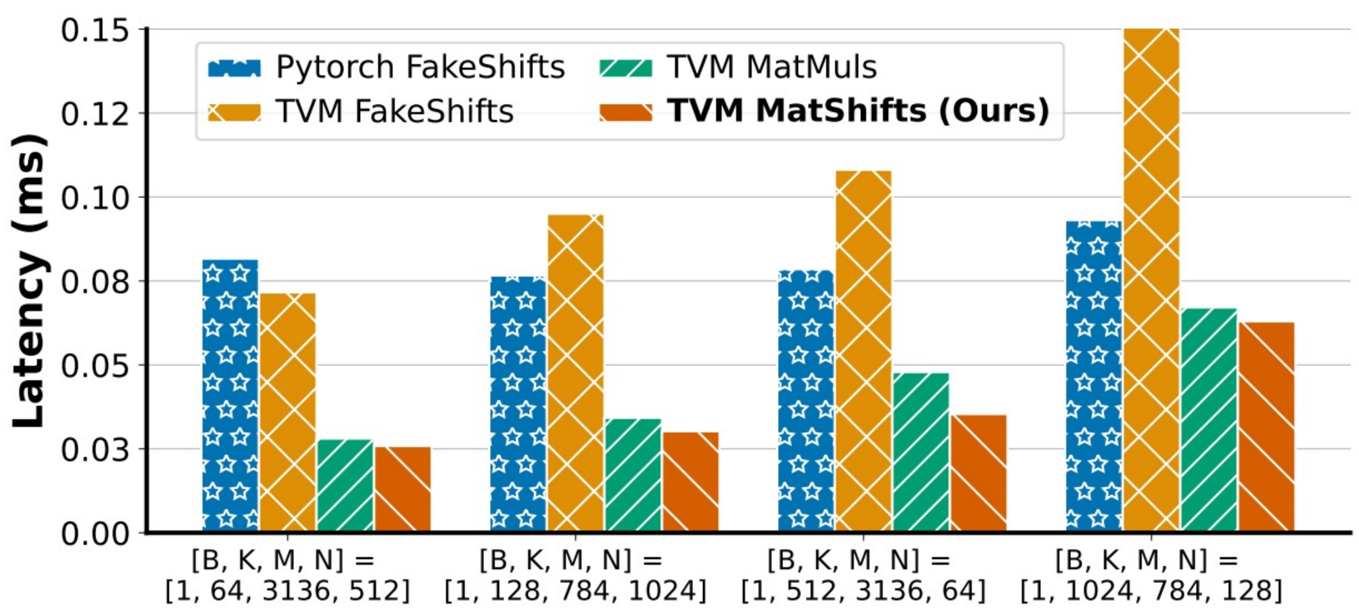
- **Overall Improvement**
  - ShiftAddViT achieves 1.74× ~ 5.18× latency reduction on GPUs and 19.4% ~ 42.9% energy savings measured on the Eyeriss accelerator with comparable or even better accuracy (↑0.04% ~ ↑0.20%)

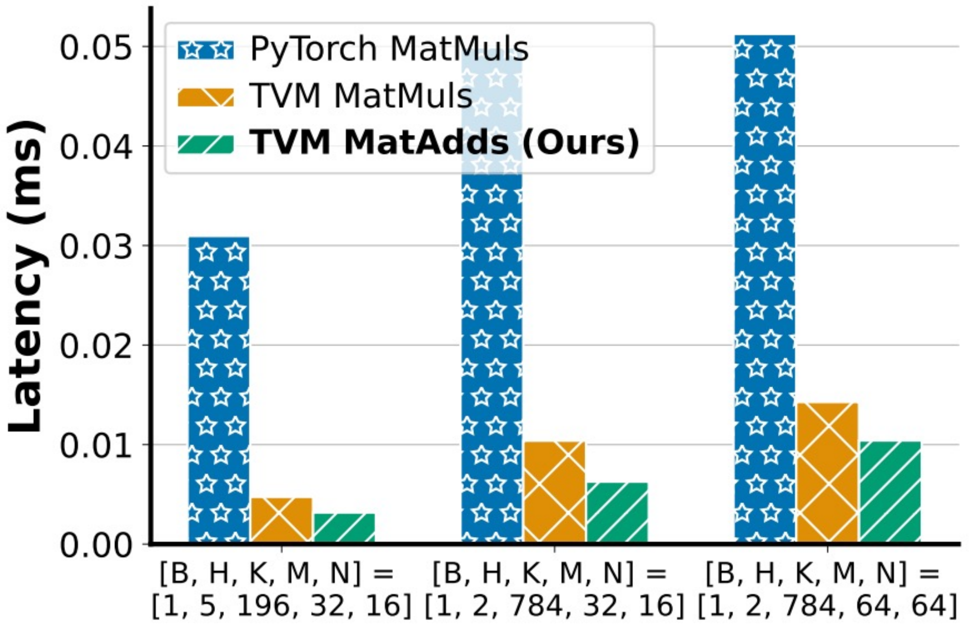
ShiftAddViT: Experimental Results for 3D Tasks

Methods	Add	Shift	MoE	LLFF Averaged			Orchids			Flower			Lat. (s)	Energy (J)
				PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS		
NeRF [41]	-	-	-	26.50	0.811	0.250	20.36	0.641	0.321	27.40	0.827	0.219	683.6	1065
GNT [53]	-	-	-	27.24	0.889	0.093	20.67	0.752	0.153	27.32	0.893	0.092	1071	1849
ShiftAddViT	✓	✗	✗	26.85	0.874	0.116	20.74	0.730	0.182	28.02	0.891	0.089	1108	1697
	✓	✓(Both)	✗	26.85	0.875	0.116	20.78	0.730	0.182	28.05	0.892	0.088	568.5	844.0
	✓	✓(Attn)	✓(MLPs)	26.92	0.876	0.114	20.73	0.731	0.180	28.20	0.894	0.087	746.6	1093
	✗	✓(Both)	✗	27.05	0.881	0.107	20.84	0.746	0.169	28.14	0.896	0.083	531.2	995.6

- **Overall Improvement**
  - ShiftAddViT achieves 22.3%/50.4% latency reductions and 20.8%/54.3% energy savings under comparable or even better generation quality (↑0.55/↓0.19 averaged PSNR across eight scenes), as compared to NeRF and GNT baselines

ShiftAddViT: Ablation Study





- **Speedups of Shifts and Adds**
  - Our MatAdds achieve on average 7.54×/1.51× speedups than PyTorch and TVM MatMuls, respectively
  - Our MatShifts achieve on average 2.35×/3.07×/1.16× speedups than PyTorch FakeShifts, TVM FakeShifts, and TVM MatMuls, respectively

Summary

For the first time, we

- **Reparameterize pre-trained ViTs with shifts and adds to deliver a new type of multiplication-reduced network, called ShiftAddViT**
- **Propose a new mixture of experts (MoE) framework for ShiftAddViT to preserve accuracy after reparameterization**
- **Conduct extensive experiments on 2D and 3D vision tasks to validate the effectiveness of our proposed ShiftAddViT**

Open-source Code:
https://github.com/GATECH-EIC/ShiftAddViT

CoCoSys
CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS